nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00957-1

Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues

Gladys Y. P. Poon[1,2] ✉, Caroline J. Watson[1,2], Daniel S. Fisher[3] and Jamie R. Blundell[1,2] ✉

Genetic alterations under positive selection in healthy tissues have implications for cancer risk. However, total levels of positive selection across the genome remain unknown. Passenger mutations are influenced by all driver mutations, regardless of type or location in the genome. Therefore, the total number of passengers can be used to estimate the total number of drivers—including unidentified drivers outside of cancer genes that are traditionally missed. Here we analyze the variant allele frequency spectrum of synonymous mutations from healthy blood and esophagus to quantify levels of missing positive selection. In blood, we find that only 30% of passengers can be explained by single-nucleotide variants in driver genes, suggesting high levels of positive selection for mutations elsewhere in the genome. In contrast, more than half of all passengers in the esophagus can be explained by just the two driver genes *NOTCH1* and *TP53*, suggesting little positive selection elsewhere.

Next-generation sequencing of healthy tissues has revealed that large numbers of nonsynonymous mutations are under positive selection and cause clonal expansions[1–12]. Expanded clones with these 'driver' mutations, which often occur in cancer-associated genes, have an increased chance of acquiring further pathogenic mutations and thus have implications for cancer risk[1,2,10,11,13,14]. Clonal expansions are typically identified via high-depth sequencing of a 'panel' of cancer-associated genes[4,5,7,8,15,16]. However, because a large fraction of the genome lies beyond the target regions of these sequencing panels and the panels are often designed to best detect single-nucleotide variants (SNVs) and indels, mutations in noncoding regions and more complex mutations including mosaic chromosomal alterations[10,11] and structural variants are potentially missed. Moreover, some cancers appear to develop without mutations affecting known driver genes, which suggests that there might be a large number of drivers that are individually rare but collectively common[6]. This raises the question: how many mutations driving clonal expansions are missed by gene-focused sequencing panels?

Positive selection for a gene or variant is commonly identified using recurrence (for example, over-representation of mutation in a particular gene)[17], elevated dN/dS ratios (the ratio of the number of nonsynonymous to synonymous mutations in a sample or gene)[4,6] and analysis of the distribution of variant allele frequencies (VAFs)[18,19]. However, none of these approaches naturally extend to quantifying the total levels of positive selection. Recurrence misses mutations that are rare[17] and confounds the effects of selection with mutation (for example, recurrence can be high at mutational hotspots)[20]. Approaches based on dN/dS are restricted to the coding sequence of genes[4,6]. Methods based on the VAF spectrum can disentangle selection from mutation and are not restricted to genes, but these also require large numbers of individuals to share the same mutation to achieve accurate estimates[18]. Therefore, estimating the total levels of positive selection outside of genes using existing methods is challenging.

Here we develop a population genetic framework that uses the distribution of VAFs of synonymous variants to quantify how much positive selection remains unexplained by mutations in canonical cancer driver genes. We show that most synonymous variants reach high VAF due to genetic hitchhiking: they are passenger mutations that co-occur with a positively selected driver mutation, which itself might be undetected. The number of these high-VAF synonymous variants thus provides information about genome-wide levels of positive selection. Specifically, each driver mutation will generate a 'comet tail' of synonymous passenger mutations as it clonally expands[19]. Once these driver mutations reach high VAF, the VAF distribution of their synonymous passengers declines with the inverse square of the VAF, often referred to as a 'Luria–Delbrück' distribution[21,22]. This characteristic distribution occurs whenever selectively neutral mutations are acquired in an exponentially growing population[23–28]. The number of passenger mutations generated in a single driver event is typically small. However, by aggregating measurements of all synonymous variants across many samples of the same tissue, a statistical distribution of synonymous passenger mutations emerges that reveals total levels of positive selection in that tissue, including positive selection on noncoding mutations and more complex genetic and epigenetic alterations.

Applying our framework to data from healthy blood[7,8,15,16], we find that high-VAF synonymous variants are predominantly passenger mutations: they exhibit the expected decline with VAF, the age dependence and the mutation co-occurrence patterns predicted by hitchhiking. However, we find that the total number of synonymous passengers is three- to fourfold higher than can be explained by nonsynonymous SNVs in canonical driver genes. This suggests high levels of positive selection elsewhere in the genome on alternative driver mutations, for example, mosaic chromosomal alterations[11]. Applying the same framework to data from healthy esophagus[5], we estimate that ~54% (range = 52–57%) of clonal expansions in healthy esophagus can be explained by mutations occurring in just two driver genes (*NOTCH1* and *TP53*) and less than ~11% (range = 4–17%) of clonal expansions remain unexplained by SNVs in known driver genes. Our method therefore suggests that there could be many noncoding, structural and copy number driver mutations under positive selection in healthy blood and relatively few in healthy esophagus. To validate this finding, we show that, in blood, samples with high-VAF synonymous variants

[1] Early Detection Programme, CRUK Cambridge Cancer Centre, University of Cambridge, Cambridge, UK. [2] Department of Oncology, University of Cambridge, Cambridge, UK. [3] Department of Applied Physics, Stanford University, Stanford, CA, USA. ✉e-mail: ypgp2@cam.ac.uk; jrb75@cam.ac.uk

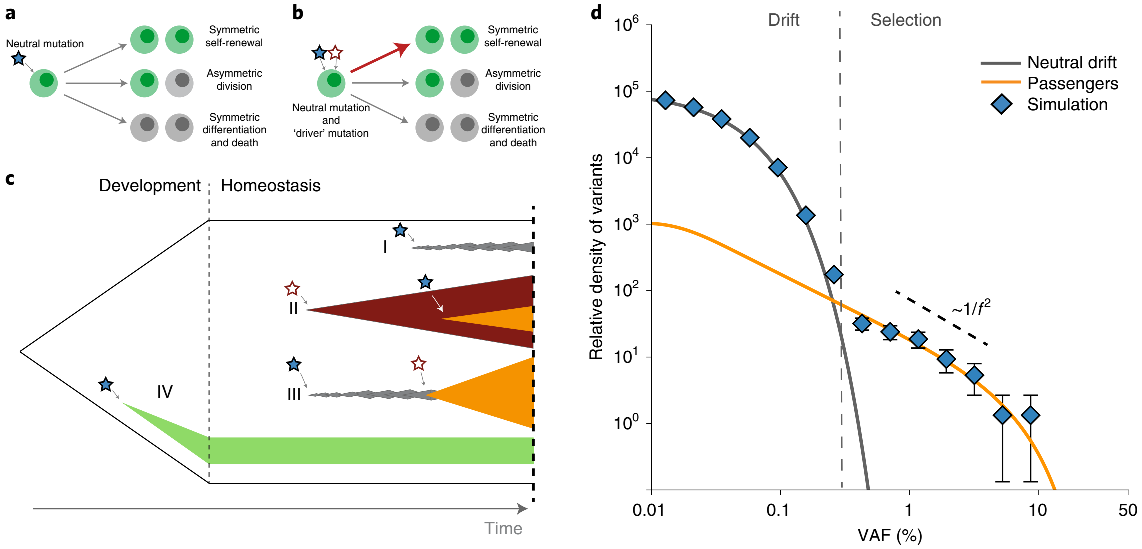


Fig. 1 | A model of genetic hitchhiking. a, In a stochastic branching model of stem cell dynamics, self-renewal and differentiation rates remain balanced in cells harboring only neutral synonymous mutations. **b**, A neutral synonymous mutation (blue star) that co-occurs with a driver mutation (maroon star) is termed a passenger. The driver causes a skew in cell fates towards self-renewal, which also impacts the passenger due to genetic linkage. **c**, A schematic depicting the four different classes of synonymous mutations that are considered here. (I) A synonymous mutation arising in adulthood not linked to any driver mutation undergoes neutral drift and remains at low VAF. (II) An expanding clone with a driver mutation subsequently acquires a synonymous passenger mutation, which hitchhikes to high VAF. (III) A neutral clone with a synonymous mutation subsequently acquires a driver mutation and hitchhikes to high VAF. (IV) Synonymous mutations that occur early during development can reach high VAF. **d**, VAF spectra of neutral mutations (blue data points) from stochastic simulations are overlaid with theoretical predictions for drift ('neutral drift', gray line) and hitchhiking ('passengers', orange line). The main figure shows an instance of simulation results (simulation run number = 10,000) (Supplementary Note 2 and Extended Data Fig. 1) for driver mutation rate $\mu_b = 3\times10^{-6}$ per year at age 70 for the case where $\tau = 1$ year. Simulated data are presented as mean values ± sampling error.

are enriched for previously unobserved driver mutations and therefore demonstrate proof of principle that our approach can guide targeted driver mutation discovery.

Results

Inferring rate of drivers from synonymous variants. To understand how the rate and selective strength of driver mutations shape the VAF distribution of synonymous variants, we first consider a stochastic model of stem cell dynamics (Fig. 1). In the initial 'development' phase, stem cells grow exponentially from a single cell until reaching a fixed population size, N. Then, in the subsequent 'homeostasis' phase, stem cell numbers remain at N by virtue of stochastically self-renewing and terminally differentiating at the same rate, $1/\tau$, where τ is the time between symmetric stem cell divisions[18] (Fig. 1a). Mutations entering the stem cell population during homeostasis are either synonymous, which are acquired at total rate μ_n, or drivers, which are acquired at total rate μ_b (both per cell per year). The rate of acquiring driver mutations is low enough that competition between two or more large driver clones within the same individual ('clonal interference') is unlikely (Supplementary Note 1b). Synonymous mutations, being neutral, do not alter the balance between self-renewal and differentiation (Fig. 1a), whereas driver mutations increase the rate of self-renewal relative to differentiation by a magnitude s per year (termed the 'fitness effect'), enabling them to exponentially expand (Fig. 1b). Fitness effects of driver mutations are drawn from a distribution of fitness effects (DFE) that reflects the range of functional consequences of different mutations. Synonymous mutations co-occurring in the same cell as a driver are termed passengers and also expand exponentially due to genetic linkage with the driver mutation. The VAF distribution of synonymous variants that results from this process has two defining features.

Synonymous variants that are detected at low VAF are most likely to have arisen during the homeostasis phase and to exist in clones that have not acquired a driver mutation (Supplementary Note 1b). These neutral clones are therefore subject to the forces of genetic drift alone (Fig. 1c, case I) and remain concentrated at low VAFs because they do not have a fitness advantage ($s = 0$). The VAF distribution that results from these synonymous variants being generated at a constant rate and subsequently drifting scales as $1/f$ at low VAF and then falls away exponentially at VAF > $\varphi = t/2N\tau$, which is proportional to age, t (refs. [18,19,29]) (Supplementary Note 1b). Because age, t, is a known quantity, the frequency at which the exponential fall-off occurs, φ, provides an important independent check on the value of $N\tau$. We validated this prediction with stochastic simulations (Supplementary Note 2) in which we recorded the density of all synonymous variants as a function of VAF, plotted on a log scale. As expected, synonymous variants at low VAF follow the distribution predicted by genetic drift alone. Their density begins to fall off exponentially at VAF > 0.03%, which, combined with the age of simulated individuals of 70 years, correctly recovers the true $N\tau = 10^5$ years used in the simulations (Fig. 1d, gray curve).

Synonymous variants detected at VAF >> φ are unlikely to be caused by drift alone. Synonymous variants reaching these higher VAFs are either passenger mutations driven to high frequency by genetic hitchhiking (Fig. 1c, cases II and III) or synonymous variants that occurred during the earliest stages of the development of the tissue (Fig. 1c, case IV). Synonymous passenger mutations can occur in two distinct ways: either a clone with a driver mutation subsequently acquires a synonymous mutation (Fig. 1c, case II) or a clone with a synonymous mutation subsequently acquires a driver mutation (Fig. 1c, case III). Because neutral clones are unlikely to survive long enough to acquire further driver mutations, synonymous mutations that subsequently acquire a driver mutation make only a small contribution to the total number of passengers and the vast majority of passengers occur by hitchhiking

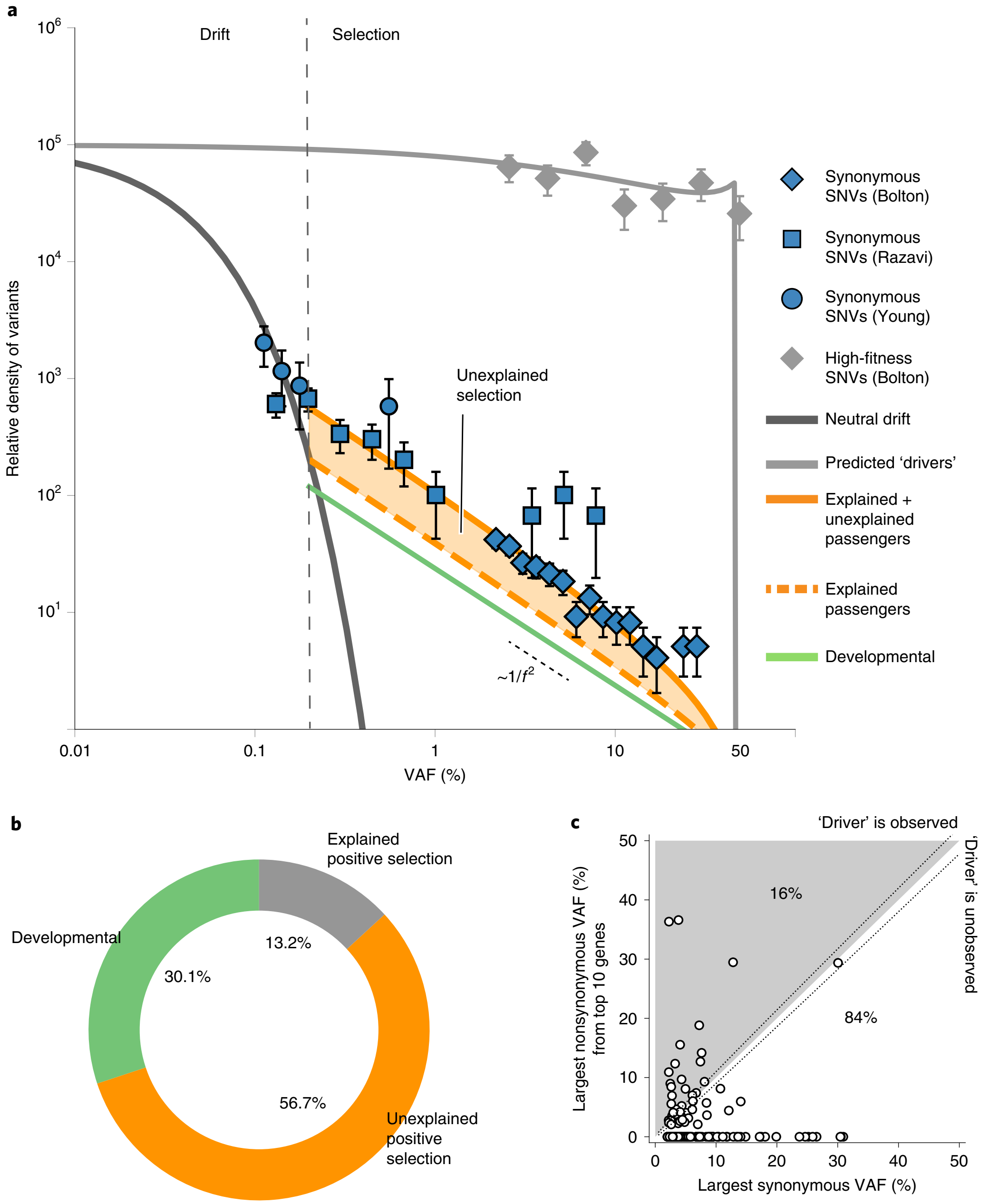


Fig. 2 | Synonymous variants in healthy blood. a, Normalized synonymous VAF spectra (blue data points, variant number $n = 344$)[15,16] are compared with passenger predictions based on driver mutations in the top 10 CH genes (orange dashed line) and predictions based on additional unobserved drivers (solid orange line). Missing positive selection is indicated by the disparity between the dashed and solid orange lines (orange shaded area). The predicted densities of neutral variants (dark gray line) and developmental mutations (green line) are highlighted. The normalized VAF spectra of the six most common high-fitness variants[15] (gray data points, variant number $n = 79$) are uniform on log(VAF), in agreement with predictions[18] (light gray line). Data are presented as mean values ± sampling error. **b**, Pie chart showing the relative contributions of developmental mutations (green), nonsynonymous drivers in the top 10 CH genes (gray) and other unobserved drivers (orange) to the total density of high-VAF synonymous mutations. **c**, Scatter plot of all synonymous variants[15] ($n = 209$) showing the VAF of the synonymous SNV and the VAF of the largest nonsynonymous SNV from the top 10 CH genes in the same individual. The diagonal dashed lines indicate the upper and lower errors due to sampling noise for exactly the same VAFs.

alongside driver mutations already in existence (Supplementary Note 1b and Supplementary Fig. 2). Mutations that occur during early development may also be observed at high VAFs and this component can be estimated using developmental mutation rates (Supplementary Note 1b).

The process of genetic hitchhiking leaves a telltale signature in the density of synonymous VAFs whereby density is expected to decline approximately with the inverse square law of VAF. This scaling can be understood by considering the case where all driver mutations confer the same fitness advantage, *s*, and where all individuals are sampled at the same age, *t*. Because driver clones expand exponentially, late hitchhiking events are exponentially more likely, but passengers arising from these late hitchhiking events will be exponentially smaller clones. This results in the density of synonymous VAFs declining with the inverse square of the VAF, a characteristic distribution that occurs whenever selectively neutral mutations are acquired in an exponentially growing population[23–28] (Supplementary Note 1a):

$$\rho(f) \approx \frac{A}{f^2} \qquad (1)$$

where

$$A = \frac{\mu_b \mu_n e^{st}}{2s^2}$$

up to a maximum VAF ~ $\varphi_s = e^{st}/2N\tau s$, the characteristic VAF of a driver clone that enters the population immediately. The density of passengers at a given VAF, controlled by A, increases with higher mutation rates to passengers (μ_n) and with the increasing expected size of the driver mutation clone. The amplitude of the inverse-square relationship thus contains information on the total levels of positive selection through the parameters μ_b and s. Because the fitness effects and mutation rates of known drivers can be estimated from the spectrum of their nonsynonymous VAFs[18], one can predict what the synonymous density would be if all hitchhiking were caused by known drivers. An observed density that is larger than can be explained by hitchhiking with known drivers implies the presence of 'missing selection' caused by unobserved drivers elsewhere in the genome. To validate this, we performed stochastic simulations that introduced synonymous mutations and nonsynonymous drivers into the stem cell pool where the fitness and mutation rates of all driver mutations were known (Supplementary Note 2). We then used the VAF spectrum of nonsynonymous mutations to infer the fitness of the drivers and inferred what the mutation rates must be to explain the amplitude observed. We were able to successfully recover the mutation rates of these known drivers, demonstrating that our framework can be used to quantify the total levels of positive selection (Fig. 1d and Extended Data Fig. 1).

Many unobserved driver mutations in healthy blood. To determine how much positive selection can be explained by known driver mutations in blood, we analyzed the VAF spectrum of ~344 synonymous variants detected in peripheral blood samples from healthy subjects from Bolton et al.[15], Razavi et al.[16] and Young et al.[7,8] (Fig. 2 and Supplementary Note 3a). We see striking agreement between all three datasets and our predictions based on a model of stem cell dynamics with genetic hitchhiking. Specifically, the density of synonymous variants declines with the inverse square of VAF (Fig. 2a, short dashed line). Over a typical human lifespan, these high-VAF synonymous variants are likely (1) developmental, (2) passengers with observed drivers or (3) passengers with unobserved drivers (Supplementary Note 1b). We outline the contributions of each below.

To determine what fraction of the high-VAF synonymous variants are developmental in origin, we inferred the developmental mutation rate for blood using data from single-cell-derived hematopoietic stem cell (HSC) colonies from Lee-Six et al.[3] and Chapman et al.[30] (Supplementary Note 3b). These datasets all point to a developmental mutation rate of 1–2 mutations per haploid genome per cell doubling and consistently show that the density of developmental mutations declines with roughly the inverse square of the VAF (Extended Data Fig. 2c). Using these estimates, we could predict the density of high-VAF synonymous variants that would be expected if all were developmental in origin (Fig. 2a, green line). Developmental mutations can explain only ~30% (range = 15–45%) of the high-VAF synonymous variants (Fig. 2b, green arc).

To estimate what fraction of high-VAF synonymous variants can be explained by hitchhiking alongside mutations in canonical clonal hematopoiesis (CH) driver genes, we used population genetic methods developed previously[18] to estimate the DFE of nonsynonymous mutations in the top 10 CH genes (ranked by number of nonsynonymous SNVs observed in Bolton et al.[15]) using all 554 observed nonsynonymous mutations in these genes (Supplementary Note 3c and Extended Data Fig. 3c). Combining this DFE with equation (15) in Supplementary Note 1, we could then predict the density of synonymous passenger mutations expected based on these observed putative drivers (Fig. 2a, orange dashed line). While the scaling of this prediction with VAF is in close agreement with the data, these observed drivers can only explain a further ~13% of high-VAF synonymous variants (Fig. 2b, gray arc). The discrepancy between the predicted density and the actual density from the synonymous variant data suggests that there is missing selection (orange shaded area, Fig. 2a). This implies that ~57% (range = 42–72%) of the high-VAF synonymous variants are not explained by development or positive selection on nonsynonymous mutations in the top 10 CH genes, that is, most positive selection occurs outside of these top 10 CH genes.

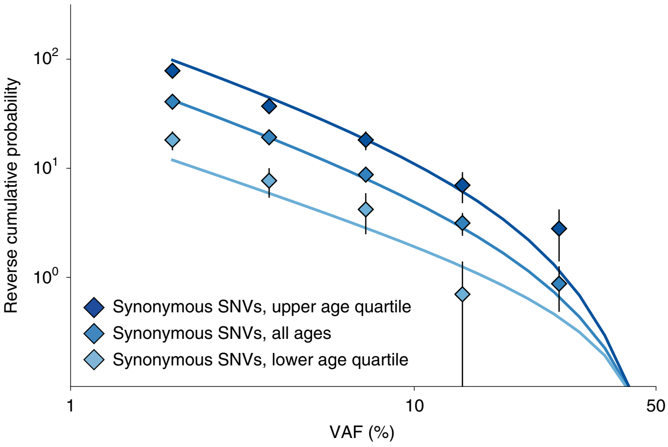


Fig. 3 | Age dependence of the synonymous VAF spectrum in healthy blood. Reverse cumulative synonymous VAF spectra for individuals in the upper quartile of ages (dark blue: age > 69 years, median 75 years, n = 1,040) and lower quartile of ages (light blue: age < 52 years, median 44 years, n = 1,040) from the Bolton cohort[15] compared with age-specific predictions (solid lines). Data are presented as mean values ± sampling error.

To explore how much of the missing selection can be explained by positive selection across a larger number of CH genes, we ranked all genes in the 468-gene panel by their number of nonsynonymous SNVs found in Bolton et al.[15] (VAF > 2%). We then estimated what fraction of passenger mutations could be explained by including progressively more genes down to lower ranks (Fig. 6, purple line). This analysis shows strong diminishing returns whereby the top 10 CH genes explain ~19% (range = 15–24%) of passenger mutations, the top 50 explain ~25% (range = 20–31%) and all 468 explain a total of only ~30% (range = 25–39%) (Supplementary Fig. 8). Thus, ~70% (range = 61–75%) of passenger mutations cannot be explained by observed nonsynonymous drivers across the 468 genes, pointing to positive selection on unobserved drivers elsewhere in the genome.

Because the density of passenger mutations expected from hitchhiking depends both on the mutation rate of drivers, μ_b, and on their fitness effects, s (equation (1)), the ~70% (range = 61–75%) of passengers that remain unaccounted for by observed nonsynonymous drivers across the 468 genes could be explained by a small number of unobserved drivers with high fitness effects or a much larger number of unobserved drivers with weak fitness effects (Supplementary Note 3f and Extended Data Fig. 4). If all unobserved drivers had fitness effects similar to the fittest CH gene mutations ($s \approx 16\%$ per year, for example, for *SRSF2*, *SF3B1*

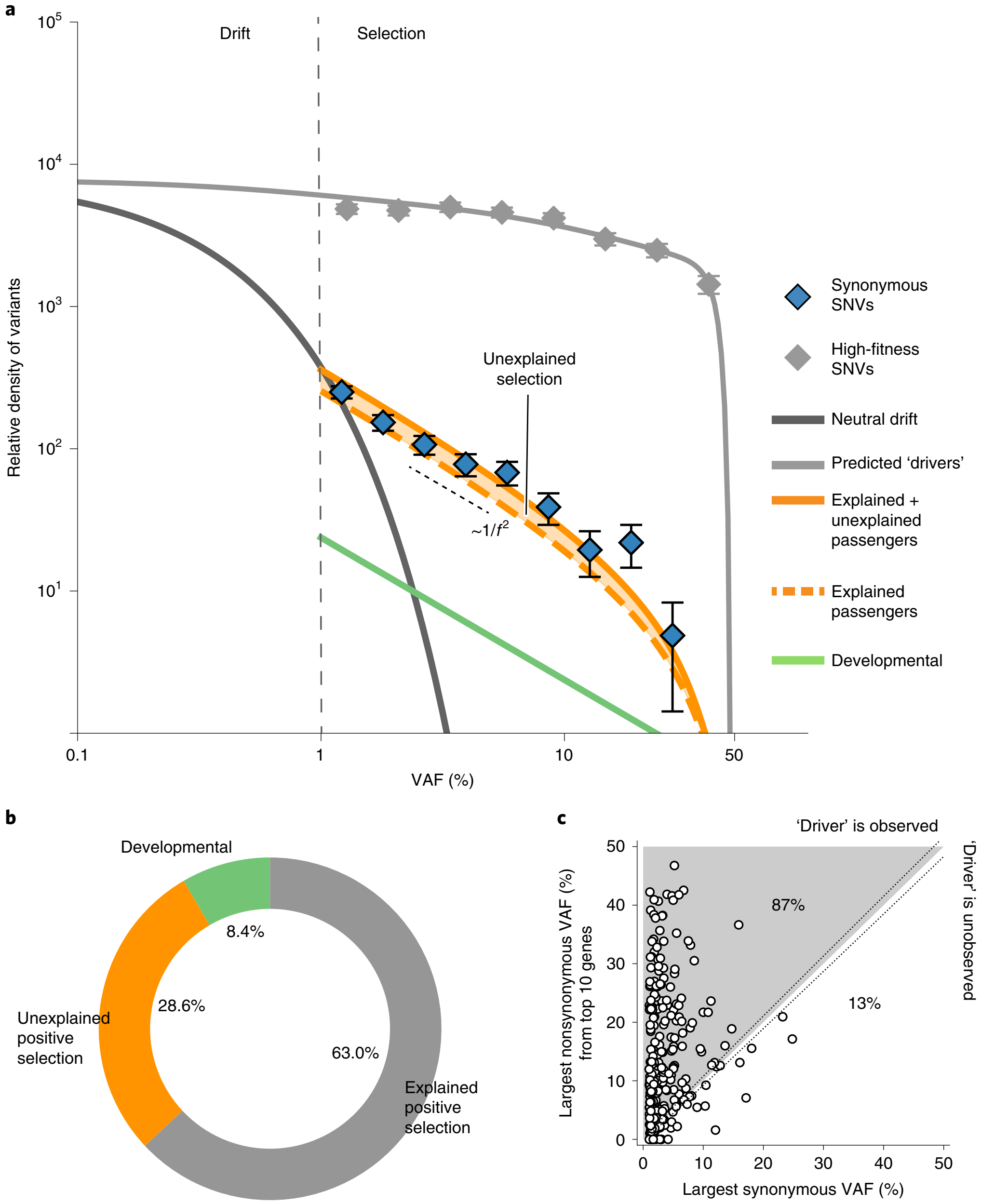


Fig. 4 | Synonymous variants in healthy esophagus. a, Normalized synonymous VAF spectra (blue data points, variant number $n = 603$)[5] are compared with passenger predictions based on driver mutations in the top 10 driver genes (orange dashed line) and predictions based on additional unobserved drivers (solid line). Missing positive selection is indicated by the disparity between the dashed and solid orange lines (shaded area). The predicted densities of neutral variants (dark gray line) and developmental mutations (green line) are highlighted. The normalized VAF spectra of the *NOTCH1* nonsynonymous variants (gray data points, variant number $n = 1{,}251$) are uniform, with log(VAF) in agreement with predictions (light gray line)[18]. Data are presented as mean values ± sampling error. **b**, Pie chart showing the relative contributions of developmental mutations (green), nonsynonymous drivers in the top 10 driver genes (gray) and other unobserved drivers (orange) to the total density of high-VAF synonymous mutations. **c**, Scatter plot of all synonymous variants[5] (individual number $n = 277$) showing the VAF of the synonymous SNV and the VAF of the largest nonsynonymous SNV in the same sample. The diagonal dashed lines indicate the upper and lower errors due to sampling noise for exactly the same VAFs.

(ref. [18])), then the missing selection could be accounted for by the equivalent of ~30 genes elsewhere in the genome. In contrast, if the unobserved drivers had fitness effects in line with the weakest detectable CH gene mutations ($s \approx 8\%$ per year (ref. [18])), the equivalent of thousands of genes elsewhere in the genome would be needed to explain the missing selection. Assuming the unobserved drivers have a similar DFE to observed drivers would mean that the rate of mutation to unobserved drivers was ~2.3-fold higher than it is to all 468 driver genes on the panel from Bolton et al.[15], a result that is broadly insensitive to different parameterized forms of the DFE (Supplementary Note 3e).

To test our finding that most selection in blood is unaccounted for by observed driver mutations, we considered co-occurrence of putative driver mutations (nonsynonymous variants in the panel) with synonymous variants within the same individual (Fig. 2c and Supplementary Note 3h). Because most hitchhiking occurs after the expansion of driver mutations (Fig. 1c, case II, and Supplementary Note 1b), the VAF of synonymous passengers cannot be much larger than the VAF of the driver mutation causing the clonal expansion (Fig. 2c, shaded region). Among individuals harboring at least one synonymous variant[15], only 16% harbor nonsynonymous variants in the top 10 CH genes whose VAF is higher than that of the largest synonymous variant (Fig. 2c, shaded region). This finding is in good quantitative agreement with our previous estimate of ~13% inferred from the fit to the synonymous VAF distribution (Fig. 2a,b). Similar agreement is seen by considering nonsynonymous mutations across a larger set of genes (Supplementary Fig. 9).

To further check the predictions of our model, we considered the age dependence of the synonymous VAF spectrum (Fig. 3). Our model predicts that the amplitude and shape of the synonymous VAF spectrum should exhibit strong age dependence, whereby high-VAF synonymous variants should be much more prevalent among older individuals (Extended Data Fig. 5). To check this prediction, we considered the lower-quartile (age < 52 years, median 44 years, $n = 1{,}040$) and upper-quartile (age > 69 years, median 75 years, $n = 1{,}040$) age groups from the Bolton cohort[15] (median = 62 years, $n = 4{,}160$) and plotted the synonymous VAF spectrum for each group against their respective predictions. The data show age dependence, in good quantitative agreement with predictions.

Few unobserved drivers in healthy esophagus. To determine how much positive selection can be explained by known driver mutations in the top 10 most commonly mutated genes in the esophagus, we analyzed the VAF distribution of ~600 synonymous variants observed in 844 2 × 2 mm^2 biopsies from healthy esophagus collected from 9 individuals[5]. In these data, the VAF is a reflection of the clonal dynamics of the self-renewing cells beneath each 2 × 2 mm^2 biopsy rather than of the entire esophagus. The VAF distribution of synonymous variants is again in close agreement with the near $1/f^2$ scaling predicted for synonymous passengers (Fig. 4a, orange line). Due to limited sensitivity at VAFs < 1%, no feature consistent with drift alone is apparent (Fig. 4a, dark gray line).

Inferences on developmental mutations are less certain in healthy esophagus, as we do not have the same kind of data where one can observe the density of developmental variants as for blood. As a reasonable starting assumption, we assume that developmental mutation rates per cell doubling are higher than somatic mutation rates per time by the same factor in any tissue (Supplementary Note 4b). Since we estimated that somatic mutation rates in esophagus are ~3-fold higher than in blood, we used a developmental mutation rate for esophagus that is 3-fold higher than the developmental mutation rate in blood. Using these estimates, we predict that only ~10% (range = 4–13%) of observed high-VAF synonymous variants are developmental in origin (Fig. 4b, green arc). While the uncertainties on this number are naturally larger than they are for blood, the contribution of developmental mutations to high-VAF synonymous variants cannot be much larger than our current estimate because, as shown below, most of the high-VAF synonymous mutation density is accounted for by hitchhiking alongside the known drivers *NOTCH1* and *TP53*.

To estimate the number of synonymous variants in healthy esophagus expected due to hitchhiking with putative driver mutations, we first estimated the fitness effects of mutations (Extended Data Fig. 6) in each of the top 10 driver genes ranked by nonsynonymous mutation count (Supplementary Note 4c). We estimate that nonsynonymous mutations in *NOTCH1* have a fitness effect of $s = 10\%$ per year and occur at a haploid rate of 4.1×10^{-5} per year, while those in *TP53* confer a fitness effect of $s = 10\%$ per year and occur at a haploid rate of 1.5×10^{-5} per year. Fitness effects of mutations in the remaining top 10 genes are estimated to be substantially smaller. Using these estimates, we could predict the density of passenger mutations explained by mutations in the top 10 driver genes in healthy esophagus (Fig. 4, orange dashed line). The scaling of this prediction with VAF is again in close agreement with the predictions of our model. The combined total amplitude of synonymous variants predicted by developmental mutations and hitchhiking alongside driver mutations in the top 10 esophagus genes, however, agrees more closely with the observed amplitude (Fig. 4a, blue data points), and only 29% (range = 24–33%) of high-VAF synonymous variants remain unexplained by either developmental mutations or mutations in one of the top 10 driver genes (Fig. 4b, orange arc).

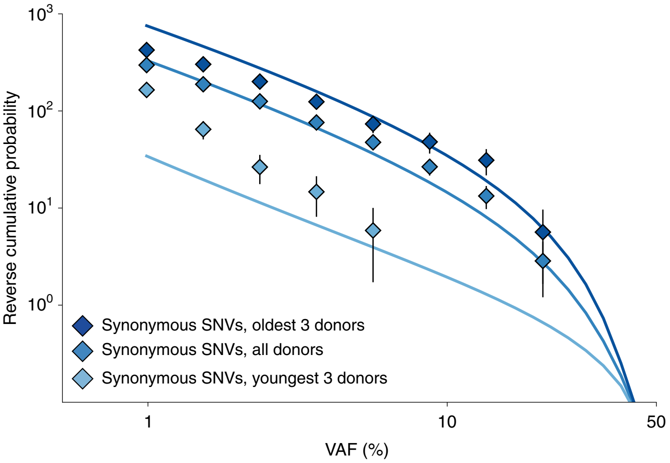


Fig. 5 | Age dependence of the synonymous VAF spectrum in healthy esophagus. Reverse cumulative synonymous VAF spectra from the youngest 3 healthy donors (light blue, age: 21.5, 25.5, 37.5 years, $n = 273$) and the oldest 3 healthy donors (dark blue, age: 57.5, 69.5, 73.5 years, $n = 284$) from Martincorena et al.[5] compared with age-specific predictions (solid lines). Data are presented as mean values ± sampling error.

To explore how much of the missing selection can be explained by positive selection across a larger number of driver genes, we again ranked genes in the 74-gene panel by their number of nonsynonymous SNVs and estimated what fraction of passenger mutations could be explained by including progressively more genes down to lower ranks (Fig. 6, green line). In stark contrast to blood, this analysis demonstrates that the top 10 driver genes explain ~69% (range = 66–72%) of high-VAF passenger mutations, the top 50 explain ~87% (range = 81–93%) and all 74 genes explain a total of ~89% (range = 83–96%). Thus, in esophagus, there is limited evidence for strong unobserved drivers elsewhere in the genome.

To test this conclusion, we considered the mutation co-occurrence of nonsynonymous driver mutations in the top 10 driver genes with synonymous variants within the same esophageal biopsy sample. The synonymous VAF distribution suggests that driver mutations in these genes account for the majority of clonal expansions; therefore, we predict that in a large fraction of biopsy samples synonymous variants will be found at a lower VAF than the largest nonsynonymous variant within the top 10 driver genes (Fig. 4c, shaded region). Consistent with this prediction, we observe that, in 87% of samples with detected synonymous variants, the largest nonsynonymous variant in the top 10 driver genes is found at a higher VAF than the largest synonymous variant (Fig. 4c and Supplementary Note 4e). This analysis implies that genetic hitchhiking in healthy esophagus is dominated by mutations in the top 10 driver genes, in fact, mostly in the top two, and that few other strong driver mutations exist elsewhere in the genome.

To further check the predictions of our model, we again considered the age dependence of the synonymous VAF spectrum. We divided the samples into a younger age group (ages = 21.5, 25.5, 37.5 years, $n = 273$) and an older age group (ages = 57.5, 69.5, 73.5 years, $n = 284$) and plotted the synonymous VAF spectrum for each group against their respective predictions. The data show clear

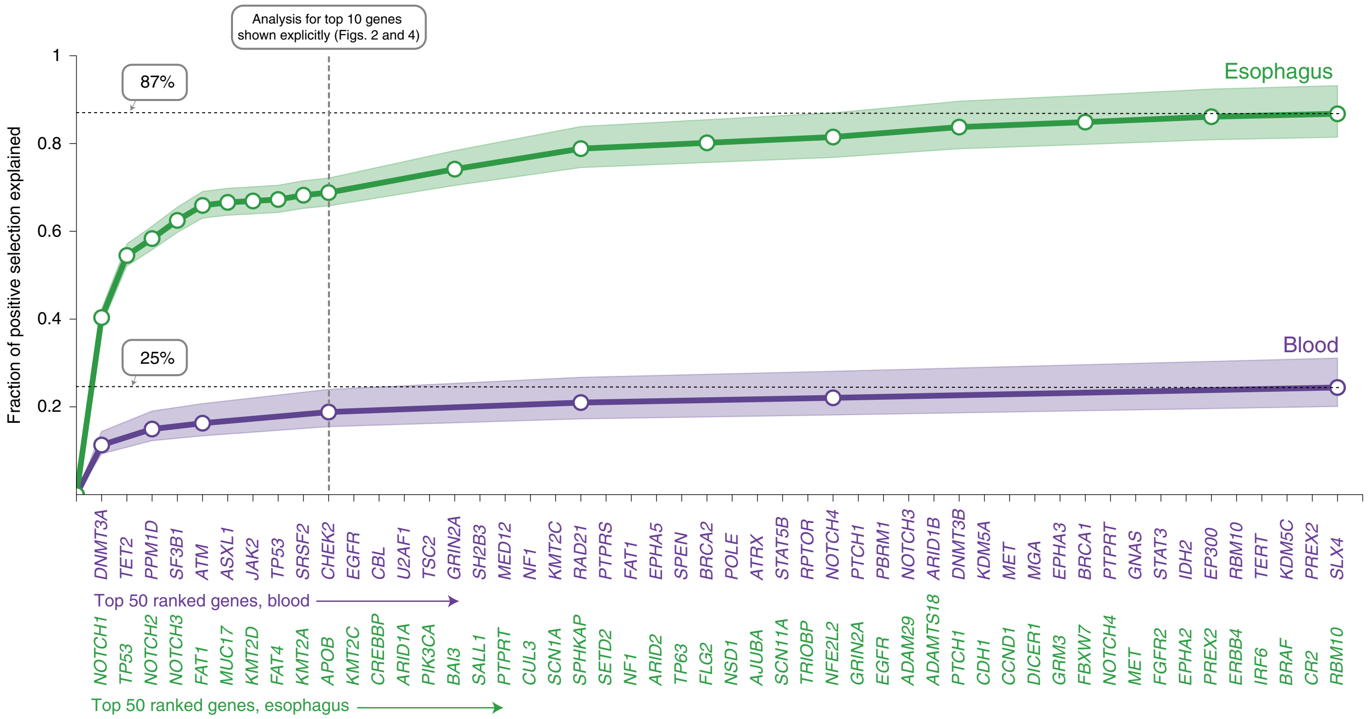
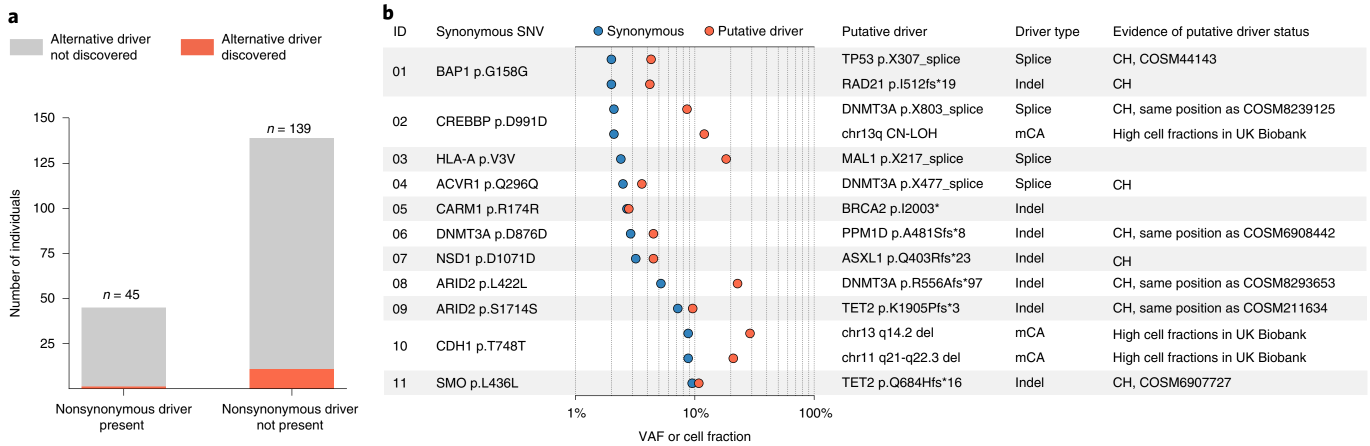


Fig. 6 | Fraction of positive selection explained by nonsynonymous variants in the top 50 driver genes in each tissue. We repeated our analysis to estimate missing selection by including increasing numbers of driver genes from 1 up to 50 (ranked by number of nonsynonymous SNVs found) in blood (purple) and esophagus (green). Ranges on the estimated fraction of explained positive selection (shaded area) were calculated for a wide range of plausible developmental mutation rates (Supplementary Notes 3b and 4b). The vertical dashed line at 10 genes highlights the data points that correspond to the analyses shown in Figs. 2 and 4.



ID	Synonymous SNV	Putative driver	Driver type	Evidence of putative driver status
01	BAP1 p.G158G	TP53 p.X307_splice	Splice	CH, COSM44143
		RAD21 p.I512fs*19	Indel	CH
02	CREBBP p.D991D	DNMT3A p.X803_splice	Splice	CH, same position as COSM8239125
		chr13q CN-LOH	mCA	High cell fractions in UK Biobank
03	HLA-A p.V3V	MAL1 p.X217_splice	Splice	
04	ACVR1 p.Q296Q	DNMT3A p.X477_splice	Splice	CH
05	CARM1 p.R174R	BRCA2 p.I2003*	Indel	
06	DNMT3A p.D876D	PPM1D p.A481Sfs*8	Indel	CH, same position as COSM6908442
07	NSD1 p.D1071D	ASXL1 p.Q403Rfs*23	Indel	CH
08	ARID2 p.L422L	DNMT3A p.R556Afs*97	Indel	CH, same position as COSM8293653
09	ARID2 p.S1714S	TET2 p.K1905Pfs*3	Indel	CH, same position as COSM211634
10	CDH1 p.T748T	chr13 q14.2 del	mCA	High cell fractions in UK Biobank
		chr11 q21-q22.3 del	mCA	High cell fractions in UK Biobank
11	SMO p.L436L	TET2 p.Q684Hfs*16	Indel	CH, COSM6907727

Fig. 7 | Targeted alternative driver discovery in individuals without nonsynonymous drivers. a, Individuals with high-VAF synonymous variants but without a nonsynonymous driver are enriched for other driver mutation types across the genome. An expanded set of mutation classes (indels, splice variants and mosaic chromosomal alterations) identifies a further 15 putative driver mutations, all of which except one occur in individuals without a high-VAF nonsynonymous variant (odds ratio = 3.9 (95% confidence interval, 0.5 to 33), $P = 0.16$, Fisher's exact test, one-sided). **b**, Table detailing the identities and VAFs (or cell fractions for mosaic chromosomal alterations (mCAs)) of the 14 alternative putative driver mutations found in individuals without a nonsynonymous driver. Variants in a gene found in CH[18] are labeled as 'CH'. Twelve of the variants show evidence of being under positive selection.

age dependence, in qualitative agreement with model predictions (Fig. 5). However, the observed age dependence is slightly weaker than predicted by the model, possibly due to an underestimation of the developmental mutation rates in healthy esophagus. This would imply that the total positive selection level is lower than our estimate and the real fraction of positive selection explained by genes in the panel is even higher.

CH driver mutation discovery. If our inferences of missing selection in blood are correct, individuals with high-VAF synonymous mutations that are not linked to a putative nonsynonymous driver would be enriched for other types of driver mutations across the genome. To demonstrate proof of principle that this could be used to guide driver mutation discovery, we obtained additional variant-class calls from individuals in the Bolton et al. cohort, including indels, splice mutations and some copy number variants[15,31]. We considered all individuals with high-VAF synonymous variants and separated them into two groups: individuals with a nonsynonymous variant at higher VAF than the synonymous variant (that is, possibly 'explaining' the passenger) and individuals without such a variant (Fig. 7a). We then identified further putative driver mutations in these new variant classes by identifying the largest clone in each of the new variant classes that is at a higher VAF than the synonymous variant detected across the individuals in both groups. We identified a total of 15 new putative driver mutations in 12 individuals across both groups, among whom all but one person belonged to the group without a putative nonsynonymous driver. Twelve of the putative alternative drivers found in individuals without a nonsynonymous driver have existing evidence of being under positive selection (Fig. 7b). Although this analysis has limitations due to the modest number of additional putative drivers (a consequence of the fact that the additional variant calls also come from a targeted panel), it nevertheless suggests that 'novel' drivers are enriched in individuals with high-VAF synonymous variants not linked with putative nonsynonymous drivers. This raises the possibility that synonymous mutations at high VAF can be used for a more targeted approach to novel driver mutation discovery.

Discussion

We have developed a population genetic framework that analyzes the VAF spectrum of synonymous variants to quantify how much positive selection in healthy tissues is explained by the known driver mutations in canonical cancer genes. The key intuition is that, because synonymous mutations are likely to have been carried to high VAF by hitchhiking with driver mutations, the number of high-VAF synonymous mutations observed across multiple samples can provide an estimate for total levels of positive selection. Comparing the observed levels of synonymous variants with the levels that would be expected by accounting for known drivers, it is possible to estimate how much positive selection is being 'missed'.

In blood, we show that the majority of synonymous variants reaching high VAF do so by virtue of hitchhiking with a driver. The large number of high-VAF synonymous variants, however, suggests that most clonal expansions are caused by drivers that fall outside of cancer-associated gene panels. Consistent with this finding, most high-VAF synonymous variants are not observed to co-occur with any candidate driver mutation. The large number of unobserved drivers probably include mosaic chromosomal alterations[10,11,32,33], large structural variants[34,35], coding mutations in non-cancer-associated genes, mutations in regulatory regions[36–39] and, possibly, epigenetic alterations[40,41]. We hypothesize that a more comprehensive survey of these alterations will yield many further mutations that will be shown to be capable of driving clonal expansions in blood[17]. Recent evidence shows that mosaic chromosomal alterations are relatively common drivers of CH[10,11] and suggests that the collective effect of all mosaic chromosomal alterations might account for a substantial proportion of the missing selection, which is an important area for future research.

In esophagus, the picture is quite different. While high-VAF synonymous variants are also consistent with being caused by hitchhiking, our analysis shows that the number of high-VAF synonymous variants is broadly what would be expected if the only mutations capable of driving clonal expansions were those in *NOTCH1* and *TP53*. In support of this, a large fraction of high-VAF synonymous variants in esophagus co-occur with a larger nonsynonymous variant in either *NOTCH1* or *TP53*. Therefore, we hypothesize that strong drivers of clonal expansions in healthy esophagus residing elsewhere in the genome are rare. One prediction made by this hypothesis is that, if esophageal squamous-cell carcinomas (ESSCs) begin with a clonal expansion in normal tissue, mutations in one of these two genes should be an early event in almost all ESSCs. Mutations in these two genes are indeed observed in ~90% of ESSCs and, at least in the case of *TP53*, are estimated to be early events in ESSCs[42,43].

Our framework for estimating how much positive selection can be explained by known cancer driver genes may be applicable to other tissues where the growth of driver mutations is unhindered by tissue structure. However, our model will require further development to be applied in settings where there is strong tissue structure (for example, colonic crypts). It is perhaps surprising that the data from the esophagus, where progenitor cells reside in a basal quasi-two-dimensional layer, are so consistent with the predictions of a model that lacks any spatial structure. Exactly how spatial structure of tissues shapes clonal dynamics is an important area of current and future research[44]. It is also important to note that our analysis of the synonymous VAF spectrum and resulting inferences of 'missing' selection depend on a number of factors, including: the inferred DFE, clonal interference, assumed neutrality of synonymous mutations, developmental mutation rates and simplifications inherent to a branching model of stem cell dynamics, each of which merit careful consideration (Methods).

The framework outlined in our work could also be used for a more targeted approach to driver mutation discovery. Because most high-VAF synonymous variants are caused by genetic hitchhiking, samples with a high-VAF synonymous variant that lack a known driver mutation will be enriched for new undiscovered drivers. In blood (and likely in other tissues too), it appears that the positive selection due to SNVs in known CH genes is only the tip of the driver mutation iceberg and that potentially many other driver mutations exist elsewhere in the genome. This is evidenced by recent studies showing the prevalence of mosaic chromosomal alterations that cause large clonal expansions in healthy blood[10,11]. As somatic evolution in various tissues becomes more comprehensively mapped, many more drivers of clonal expansions will emerge, leading to better understanding of cancer evolution that can inform treatment decisions and point toward potential therapeutic targets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00957-1.

Received: 8 October 2020; Accepted: 20 September 2021;
Published online: 4 November 2021

References

1. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
2. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
3. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
4. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
5. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
6. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
7. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, 12484 (2016).
8. Young, A. L., Tong, R. S., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis and risk of acute myeloid leukemia. *Haematologica* https://doi.org/10.3324/haematol.2018.215269 (2019).
9. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
10. Loh, P.-R. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).

11. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* https://doi.org/10.1038/s41586-020-2430-6 (2020).
12. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
13. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
14. Desai, P. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med.* **24**, 1015–1023 (2018).
15. Bolton, K. L. et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nat. Genet.* https://doi.org/10.1038/s41588-020-00710-0 (2020).
16. Razavi, P. et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat. Med.* **25**, 1928–1937 (2019).
17. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
18. Watson, C. J. et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).
19. Williams, M. J. et al. Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dN/dS ratios. *eLife* **9**, e48714 (2020).
20. Hess, J. M. et al. Passenger hotspot mutations in cancer. Preprint at *bioRxiv* https://doi.org/10.1101/675801 (2019).
21. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511 (1943).
22. Desai, M. M. & Fisher, D. S. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759–1798 (2007).
23. Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
24. Loeb, L. A. et al. Extensive subclonal mutational diversity in human colorectal cancer and its significance. *Proc. Natl Acad. Sci. USA* **116**, 26863–26872 (2019).
25. Blundell, J. R. et al. The dynamics of adaptive genetic diversity during the early stages of clonal evolution. *Nat. Ecol. Evol.* **3**, 293–301 (2019).
26. Fusco, D., Gralka, M., Kayser, J., Anderson, A. & Hallatschek, O. Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments. *Nat. Commun.* **7**, 12760 (2016).
27. Schreck, C. F. et al. Impact of crowding on the diversity of expanding populations. Preprint at *bioRxiv* https://doi.org/10.1101/743534 (2019).
28. Lohmueller, K. E. et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet.* **7**, e1002326 (2011).
29. Simons, B. D. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proc. Natl Acad. Sci. USA* **113**, 128–133 (2016).
30. Chapman, M. S. et al. Lineage tracing of human embryonic development and foetal haematopoiesis through somatic mutations. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.29.088765 (2020).
31. Gao, T. et al. Interplay between chromosomal alterations and gene mutations shapes the evolutionary trajectory of clonal hematopoiesis. *Nat. Commun.* **12**, 338 (2021).
32. Danielsson, M. et al. Longitudinal changes in the frequency of mosaic chromosome Y loss in peripheral blood cells of aging men varies profoundly between individuals. *Eur. J. Hum. Genet.* **28**, 349–357 (2020).
33. Thompson, D. J. et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
34. Miyamoto, T., Weissman, I. L. & Akashi, K. AML1/ETO-expressing nonleukemic stem cells in acute myelogenous leukemia with 8;21 chromosomal translocation. *Proc. Natl Acad. Sci. USA* **97**, 7521–7526 (2000).
35. Corces-Zimmerman, M. R. & Majeti, R. Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis. *Leukemia* **28**, 2276–2282 (2014).
36. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
37. Khurana, E. et al. Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17**, 93–108 (2016).
38. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
39. Kumar, S. et al. Passenger mutations in more than 2,500 cancer genomes: overall molecular functional impact and consequences. *Cell* https://doi.org/10.1016/j.cell.2020.01.032 (2020).
40. Li, S. et al. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat. Med.* **22**, 792–799 (2016).
41. Gebhard, C. et al. Profiling of aberrant DNA methylation in acute myeloid leukemia reveals subclasses of CG-rich regions with epigenetic or genetic association. *Leukemia* **33**, 26–36 (2019).
42. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
43. Liu, X. et al. Genetic alterations in esophageal tissues from squamous dysplasia to carcinoma. *Gastroenterology* **153**, 166–177 (2017).
44. Colom, B. et al. Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. *Nat. Genet.* https://doi.org/10.1038/s41588-020-0624-3 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Blood datasets. The principal dataset used to analyze the VAF spectrum of synonymous variants in cancer-free blood is from Bolton et al.[15]. This dataset can be downloaded using the link https://raw.githubusercontent.com/papaemmelab/bolton_NG_CH/master/M_long.txt.

The table from the link above includes all 11,076 mutation calls made from 24,146 individuals across the Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) panel, covering 341 (v.3), 410 (v.5) and 468 (v.6) cancer-associated genes. The associated sizes of the coding regions for each panel version are 0.9 Mb (v.3), 1.0 Mb (v.5) and 1.1 Mb (v.6). The table also details cancer and treatment status.

To examine the VAF spectrum in 'healthy' blood, individuals with hematological malignancies were excluded in the original study and we also further excluded any individuals who had received cancer treatment, as this is known to affect selection pressures[15]. This left 4,160 untreated patients (who had a nonhematologic cancer). The median sequencing depth in these data was 665×, enabling high-confidence SNV calls down to 2% VAF, which resulted in 233 synonymous and 1,236 nonsynonymous SNVs. False-positive and false-negative rates are expected to be low for these SNVs. A 'panel of normals' from 300 individuals <20 years old was used to remove sequencing artifacts. The recall rate from replicate sequencing preps for SNVs identified at VAF >2% was reported at >90% and showed a correlation coefficient of 0.98 (see Bolton et al. methods, 'Validation of calls'[15]).

To validate the VAF spectrum, we checked to see whether it was consistent with data reported by Razavi et al.[16], which can be downloaded from the European Genome-Phenome Archive (EGA) under accession no. EGAS00001003755. Synonymous variants were originally filtered out by Razavi et al.[16]. However, after consulting with the authors, we were able to call synonymous SNVs by using the same mutation calling pipeline, which can be downloaded here: https://github.com/ndbrown6/MSK-GRAIL-TECHVAL/blob/master/R/0_exec_05_chip_mutation_wbc.R, and removing the filters in lines nos. 368–369, 387–398 and 409–410, which filtered out synonymous variants.

We examined the resulting 81 synonymous SNVs and 302 nonsynonymous SNVs called from the white blood cells of the 47 healthy controls. This study adopted an error-correctable duplex sequencing approach with a minimum raw average target depth of 60,000× on a panel with coding regions spanning 1.3 Mb. This sequencing approach relies on sequencing copies of the same template molecules many times to suppress errors. Mutation calls were then filtered by a number of downstream steps to remove technical artifacts and residual noise[16]. The average template depth of the resulting mutation calls from the white blood cells of 47 healthy controls was ~3,800× with a minimum of 2 consensus molecules needed to support the variant. As a result of the increased depth and error-correctable strategy, this study was able to report variants at lower VAFs compared with Bolton et al. As reported in the original study, the reproducibility of mutation calls was high even at VAFs <1% (Razavi et al.[16], Fig. 1d–f). Even though the original study reported variant calls down to <0.1%, we excluded SNVs with VAF <0.1% because the density of SNVs declines below this value, indicating an increasing false-negative rate. The close agreement between the densities inferred from Bolton et al. and Razavi et al., despite very different sequencing approaches, indicates that the majority of these calls are true somatic variants and independently validates the densities estimated from Bolton et al.[15].

To further validate the VAF spectrum, we also compared the predictions with 30 synonymous variants from 89 cancer-free individuals from Young et al. (2016)[7] and Young et al. (2019)[8]. These sequencing datasets were generated using the Illumina TruSight Myeloid Sequencing panel (panel size of 141 kb) and also adopted an error-correctable sequencing approach with unique molecule identifiers to label single strands. While the original studies reported detection limits down to 0.03% VAF, we excluded any SNVs reported below 0.1% VAF because of the risk of false negatives. The original study, however, reported very good VAF correlation ($R^2 = 0.98$) with digital droplet PCR validations even at frequencies <0.5%, suggesting modest false-negative rates even at low VAF.

Esophagus dataset. The sequencing data for healthy esophagus were originally reported by Martincorena et al.[5]; they may be found in the EGA under accession codes EGAD00001004158 and EGAD00001004159 and can be downloaded directly here: https://www.science.org/doi/suppl/10.1126/science.aau3879/suppl_file/aau3879_tables2.xlsx.

This dataset was generated from 844 2×2 mm^2 biopsies from postmortem esophageal squamous epithelium from 9 donors. Each sample was sequenced on a targeted panel comprising 74 cancer-associated genes, spanning 330 kb, to an average depth of 870×. Each of the 844 biopsies was treated as a sample. A total of 603 synonymous and 4,371 nonsynonymous mutations were reported. As outlined in the supplementary materials of Martincorena et al.[5] (section 3, 'Mutation calling'), mutations identified showed excellent agreement between independent replicate DNA preps, with >95% of the mutations being detectable in the second replicate. False-positive rates are expected to be low as the original study utilized a panel of normals to eliminate artifacts. As always, false-negative rates towards the limit of detection of the assay could be substantial. To minimize the risk of false negatives at low VAF, we restricted to considering SNVs with >1%, which is ~10-fold the reported detection limit.

Variants detected across multiple biopsies. We analyzed all variant calls after trimming, including collapsed variant calls spanning multiple biopsy samples. Within the same biopsy, mutations shared between samples closer than 10 mm were collapsed in the original study[5]. For example, the same exact variant with VAF of 5% in one tile and 10% in an adjacent tile would be counted as one clone of VAF 15%. Because the design of the esophagus study was to tile the biopsy with the 2×2 mm^2 samples, this approach means that VAF estimates from merged variants more closely reflect the true clone size. Had we used only unmerged variants, any clones spanning boundaries would have clone sizes underestimated. More than 30% of *NOTCH1* (415 of 1,251) and *TP53* (145 of 451) nonsynonymous calls resulted from merging as described above, whereas only a small portion (57 of 603, <10%) of synonymous variant calls resulted from merging because VAFs associated with synonymous variants are typically smaller. Our analysis of passenger VAF spectra in the merged set is therefore less limited by finite biopsy sample sizes.

Recovery of driver mutation rates and selection levels in simulations. We used stochastic simulations to validate our predictions and test the robustness of our model.

Our stochastic model is based on a continuous-time branching process[18,22] of cell births and deaths whereby the probability of a stem cell symmetrically producing two terminally differentiated cells (equivalent to dying) is $D\delta t$ and the probability of a stem cell dividing symmetrically, producing two stem cells, is $B\delta t$ over a small time interval δt (B and D represent birth and death rates, respectively). In the stochastic simulations, we set $\delta t = 0.1$ in units of generations (that is, time over which $D = 1$) as an approximation to a continuous-time process and modeled the number of births and deaths as Poisson-distributed around their expected values.

Mutation fitness effects were incorporated in birth rates $B = 1 + s_{\mathrm{rel}}\tau$, where s_{rel} is the relative fitness to the mean fitness of the population per year, which typically grows in time as more and more driver mutations are introduced into the population, and τ is the generation time measured in years. The probability of producing a new mutant cell in the next generation from an existing clone of size n is always un, where u represents the mutation rate per cell per generation. We assume fitness effects of new mutations are additive, such that a clone with two mutations with fitness effects s_1 and s_2 in a wild-type population would have a fitness advantage of $s_1 + s_2$. The simulated neutral VAF spectra match our predictions closely for neutral drift and passengers (Supplementary Figs. 1 and 2).

In the simulations, cell fates are stochastic and the population size fluctuates due to drift. Changes in the wild-type stem cell population size due to drift are controlled by the parameter combination $t/N\tau$, which is small over a human lifespan for the estimated $N\tau$ parameters. Therefore, stochastic fluctuations in the size of the stem cell pool can be neglected.

As more driver mutations are introduced to the population, with these drivers expanding exponentially, the total pool of stem cells will also eventually expand exponentially—which manifests in the mean fitness of the population increasing. At this point, one has to decide whether there is clonal interference (competition between clones due to a fixed population of stem cells) or whether one allows the total population size to increase exponentially. The difference between these two implementations becomes important when mutant clones compose a large proportion of the population. However, it is important to clarify that, for the clonal dynamics of blood and esophagus, the average mutant proportion is typically small and therefore whether one implements clonal interference via a changing mean fitness or not typically has little impact on the results.

We also set up simulations to simulate the developmental stage, whereby the stem cell population grows deterministically at an exponential rate, during which mutations occur according to the developmental mutation rate. All mutations behave as neutral (even if they are assigned as beneficial) and expand deterministically until the stem cell population reaches the adult population size, at which point stem cells' fates become stochastic and cells harboring beneficial drivers produce more offspring than cells with completely neutral mutations. Our theoretical prediction (Supplementary Note 1, equation (6)) matches with the simulated VAF spectra of neutral mutations that arise during development (Supplementary Fig. 3, green diamonds).

We tested the ability of our model to accurately recover the total driver mutation rate and the driver mutation fitness across a range of mutation rates. First, a two-parameter maximum-likelihood fit for fitness effect and mutation rate was performed on the 'nonsynonymous' VAF spectrum generated in simulations, following the same procedure as outlined in Watson et al.[18] and using equation (16) in Supplementary Note 1. Second, the inferred fitness effect was then fixed at its maximum-likelihood value and a one-parameter fit was then performed on the 'synonymous' VAF spectrum for the total driver mutation rate.

The likelihood optimization was performed using the 'Nelder–Mead' algorithm with respect to reverse cumulative densities across bins in log scale above the observed drift limit $\Psi \approx 3 \times 10^{-3}$ (the VAF at which the abundance of mutants driven by drift becomes less than that of hitchhiking mutants) and below the maximum VAF any variant attains (because the detectability of variants then becomes limited by the number of simulation runs). Sampling errors were modeled as Poisson, for which the standard error of the SNV count in a bin containing m SNVs is $\sqrt{m}$; that is, the fractional error of the SNV count in bins with more SNVs is smaller. They were weighted during the fitting process so that data points with larger errors contributed less.

This approach was able to accurately recover total driver mutation rates (Extended Data Fig. 1). The modest but systematic underestimation of the total rate for high driver mutation rates was due to clonal interference between clones within an individual. This effect is not considered here because clonal interference between two or more high-VAF clones is uncommon in both the blood and the esophagus data. However, it becomes important when $N\tau\mu > 1$.

Accounting for a range of ages and fitnesses. To apply our predictions to real, rather than simulated, data, we must account for two important complications. First, driver mutations will confer a range of different fitness effects. Second, individuals have a range of ages at time of sampling. We therefore extended equation (1) (main text) to account for a DFE and a distribution of ages (Supplementary Note 1b). With these additional factors, the expression for the density of synonymous variants from equation (1) becomes more cumbersome and there is no longer a universal scaling with VAF due to the range of s and ages (Supplementary Note 3d). Nevertheless, the key features of the predicted distribution can be understood by considering which values of s contribute most to the passenger spectrum at different ages (Extended Data Fig. 5). This is determined both by the DFE and by the age of the individual. For example, at early ages, drivers with small fitness effects contribute substantially to the passenger spectrum due to their higher rate of occurrence, whereas at late ages drivers with large fitness effects, despite being rarer, contribute most to the passenger spectrum because of their exponentially larger clone sizes. Thus, even with a range of different fitness effects and a range of ages, the amplitude of high-VAF synonymous variants can be predicted using our framework.

Framework assumptions. *Sensitivity to fitness effects.* The amplitude of high-VAF synonymous variants depends exponentially on the fitness effects of linked driver mutations. Therefore, uncertainties in the DFE of driver mutations can lead to large uncertainties in the estimates for how much of the positive selection is explained. To investigate the robustness of our inferences to these effects, we explored a number of different possible forms for the DFE (Supplementary Note 3e). While the details of the functional forms for the DFE naturally make small quantitative differences to our results, they do not alter the broad qualitative conclusions. The sensitivity to fitness effects also impacts the interpretation of missing selection: it is clear from the data that there are too many high-VAF synonymous SNVs in blood to be explained by mutations in the top CH genes, and this points to positive selection on mutations outside of these genes. Whether the missing selection is made up of large numbers of smaller-effect mutations or a small number of larger-effect mutations, however, is hard to tease apart from the synonymous spectrum alone and will require further direct investigation.

Clonal interference and multiple mutant clones. Our predictions for the VAF distribution of passengers assume that clonal interference—competition between independent driver clones in the same individual or biopsy—is rare. Data from blood suggest that this is a good assumption: in healthy individuals with driver clones in blood, there is typically only one large driver clone. The parameter combination that controls clonal interference also controls the prevalence of clones with multiple driver mutations, and therefore multiple mutant clones are also expected to be rare in blood (Supplementary Note 1b). In esophagus, the data show that in about a quarter of the biopsies there is evidence of clonal interference (Supplementary Fig. 12), an observation consistent with recent studies showing clonal collision[44]. It is therefore possible that in these data there is the presence of multiple mutant clones. To check what impact this could have on our inferences, we simulated clonal dynamics with parameters inferred from the esophageal data both with and without clonal interference implemented for a range of fitness effects (Supplementary Fig. 13). Implementing clonal interference via a fixed population size or simply measuring clone VAFs by normalizing by the total population of cells makes very little quantitative difference to the VAF spectrum of single mutants. Therefore, clonal interference between single mutants does not impact our analysis. Clonal competition, however, would cause differences in the VAF spectrum at later times as the difference between implementing clonal competition via a fixed population size versus measuring sizes relative to an increasing population size begins to be notable once single mutants constitute a large fraction of the population[22].

Synonymous mutations subject to selection. An important assumption in our analysis is that synonymous variants are selectively neutral during somatic evolution. Evidence from cancer genomes suggests that some synonymous variants have functional consequences[45,46], possibly due to codon usage bias[47,48]. However, it is likely that these variants constitute a small minority of all synonymous variants and confer weak fitness effects[39] (Supplementary Tables 1–3).

Developmental mutations. Synonymous mutations that occur during development can explain the presence of some high-VAF synonymous mutations. The accuracy of our estimate for missing positive selection in healthy tissues therefore relies on the accurate quantification of developmental contribution in these tissues. For blood, whole-genome sequencing of many single-HSC-derived colonies from the same individual can be used to build a phylogenetic tree for HSCs that can be used to confirm which mutations arise during development and thus can be used to estimate developmental mutation rates[3]. It will be important to check the robustness of these inferences in future studies by directly analyzing VAF distributions in large cohorts of younger individuals.

A simplified branching model. We have used an intentionally simplified model of stem cell dynamics in which many complicating factors, such as aging, the microenvironment and spatial effects, are not explicitly accounted for. While these effects likely do influence stem cell dynamics, in previous work we have shown that this simple stochastic branching model can explain many of the quantitative aspects of the VAF data and is robust to many complications added to the model[18]. Moreover, two features in the data here suggest that this simple model captures much of the relevant behavior. First, the data clearly show that variants under strong positive selection have a different VAF dependence from passenger mutations, scaling as $1/f$ versus $\sim 1/f^2$ as the model predicts. Second, the high-depth blood sequencing data suggest that as VAFs drop below ~0.2% there may be the beginnings of a transition from selection-dominated behavior to drift-dominated behavior. While this requires deeper blood sequencing data to ascertain, a transition at this VAF quantitatively agrees with the predictions of this model and with previous estimates of the key parameter $N\tau$ (refs. [3,18]). This latter point demonstrates the value of studies with highly sensitive sequencing capable of detecting low-VAF neutral clones[7,8,16], as they provide important internal consistency checks on the model.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The principal dataset from Bolton et al. can be downloaded using the link https://raw.githubusercontent.com/papaemmelab/bolton_NG_CH/master/M_long.txt. The dataset from Razavi et al. can be downloaded from the European Genome-Phenome Archive (EGA) under accession no. EGAS00001003755. All synonymous variants analyzed in this manuscript are listed in Supplementary Tables 1–3. The sequencing data for healthy esophagus were originally reported by Martincorena et al.; they may be found in the EGA under accession codes EGAD00001004158 and EGAD00001004159 and can be downloaded directly from https://www.science.org/doi/suppl/10.1126/science.aau3879/suppl_file/aau3879_tables2.xlsx.

Code availability

All code used in this study will be available on the Blundell laboratory GitHub page: https://github.com/blundelllab/Genetic-hitchhiking.

References

45. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324–1335 (2014).
46. Sharma, Y. et al. A pan-cancer analysis of synonymous mutations. *Nat. Commun.* **10**, 2569 (2019).
47. Supek, F., Skunca, N., Repar, J., Vlahovicek, K. & Smuc, T. Translational selection is ubiquitous in prokaryotes. *PLoS Genet.* **6**, e1001004 (2010).
48. Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).

Acknowledgements

We thank K. Bolton, A. Zehir and E. Papaemmanuil for sharing unpublished data. We also thank D. Solit, P. Razavi, D. Brown and J. Reis-Filho for sharing data and I. Martincorena for sharing data and for discussions. G.Y.P.P., C.J.W. and J.R.B. are funded by the CRUK Cambridge Centre and CRUK Early Detection Programme. J.R.B. is supported by a UKRI Future Leaders Fellowship. D.S.F. and J.R.B. are supported by the Stand Up to Cancer Foundation and the National Science Foundation via grant no. PHY-1545840.

Author contributions

J.R.B. conceived the project. G.Y.P.P. developed the theory with input from J.R.B. and D.S.F. Data analysis methods, plotting and numerical simulations were all developed by G.Y.P.P. with input from J.R.B. and C.J.W. The manuscript was written by G.Y.P.P. and J.R.B., with input from C.J.W. All authors provided comments and edits on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

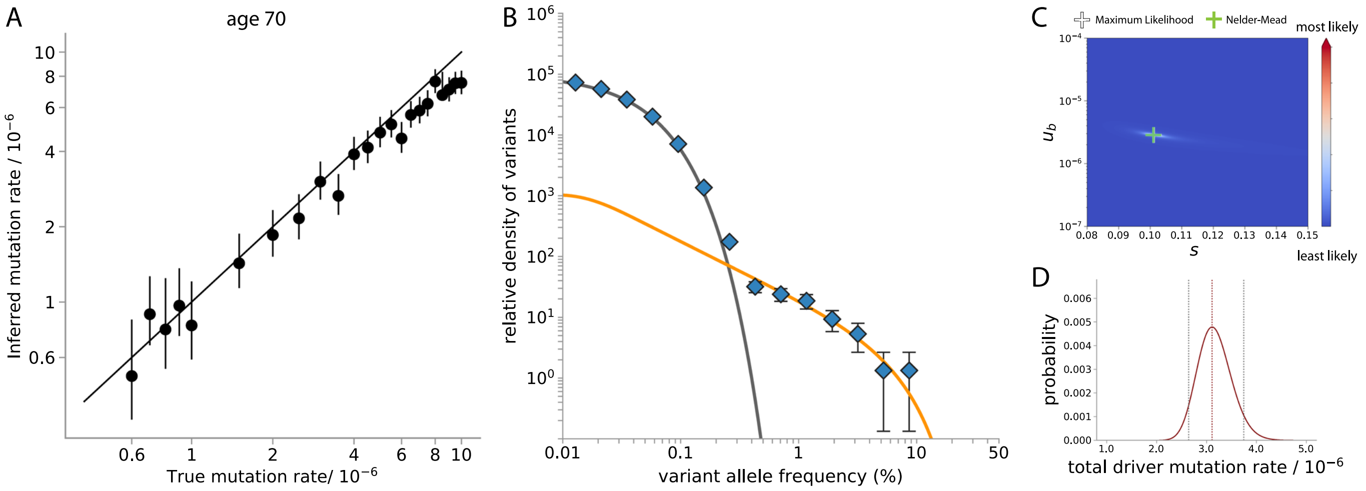
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00957-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00957-1.

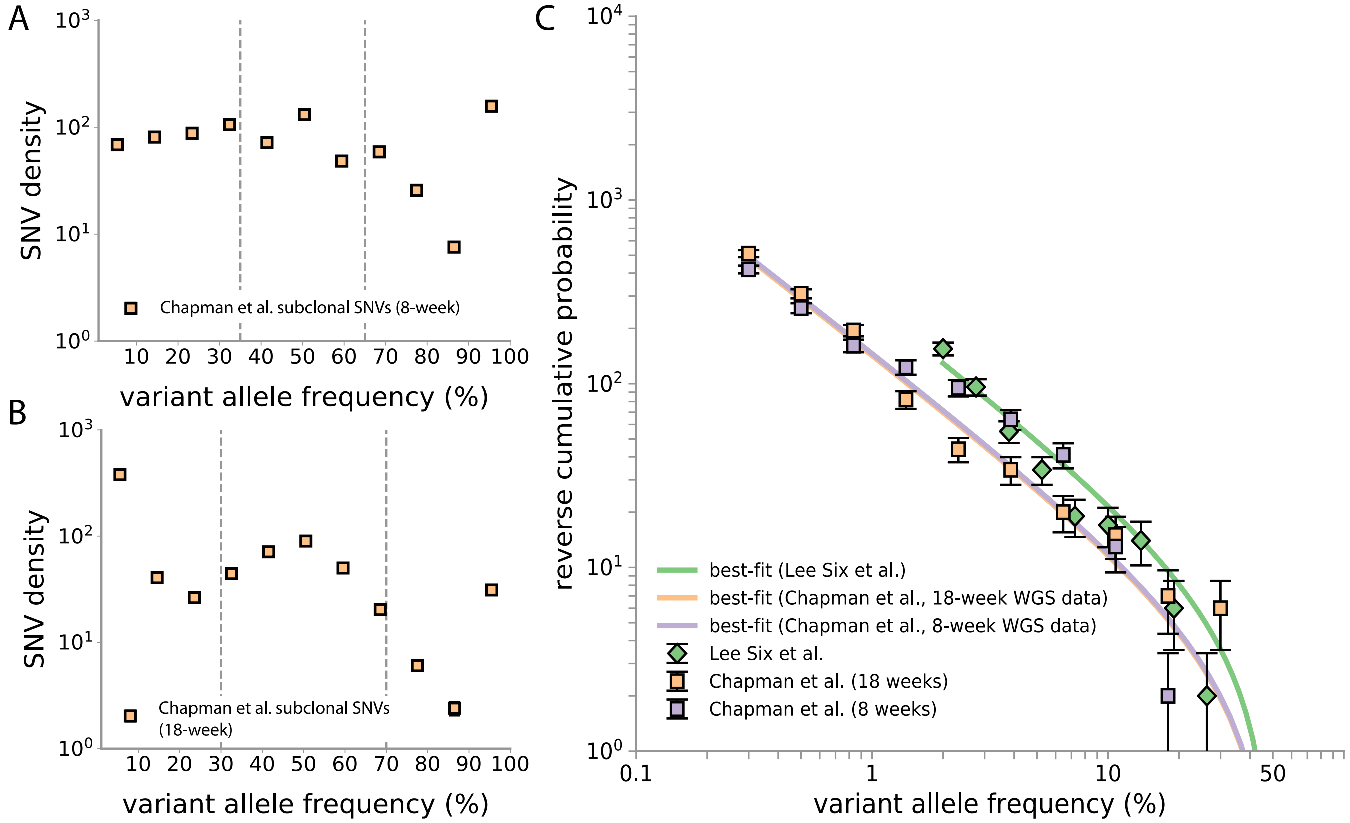
Correspondence and requests for materials should be addressed to Gladys Y. P. Poon or Jamie R. Blundell.

Peer review information *Nature Genetics* thanks Ruben van Boxtel, Benjamin Werner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

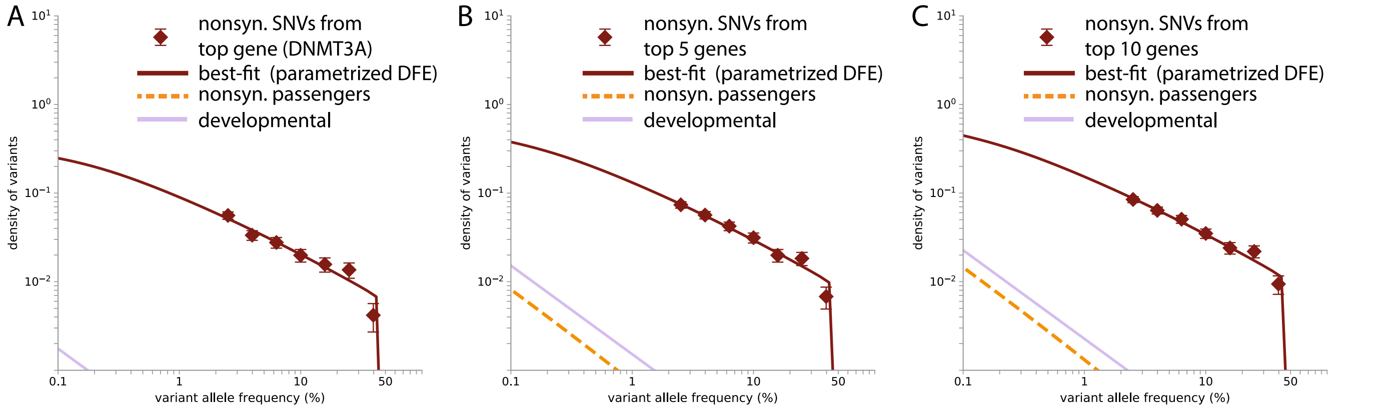
Reprints and permissions information is available at www.nature.com/reprints.



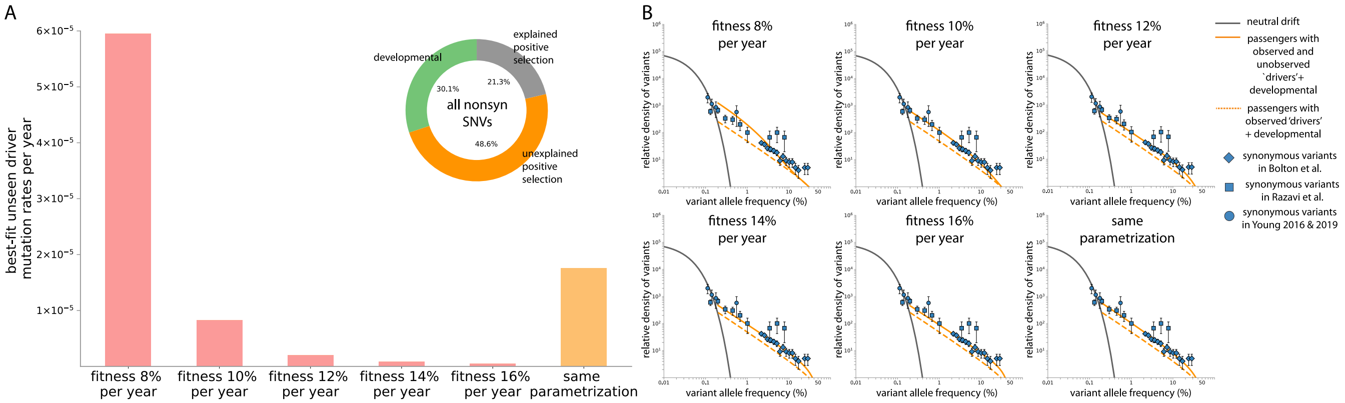
Extended Data Fig. 1 | Model performance in recovering driver mutation rates in simulations. (**a**) Our method is able to recover driver mutation rates accurately across a range of mutation rates (5×10^5 simulation runs were performed). At higher driver mutation rate, it is mainly limited by clonal interference which causes clones to reach sizes lower than that predicted by our theory. Best-fit values are presented with their 95% confidence intervals. (**b**) This shows the simulation (run no. = 15000) corresponding to driver mutation rate $\mu_b = 3\times10^{-6}$ ($\tau = 1$ year). The neutral mutation frequency spectrum above $\Psi = 3\times10^{-3}$ was fitted with our passenger prediction to infer the underlying driver mutation rates driving the expansions. Simulated data are presented as mean values ± sampling error. (**c**) The likelihood plot shows the fit for the driver mutation rate and fitness by examining the 'nonsynonymous' variant allele (that is driver mutation) frequency spectrum only. It is overlaid with the maximum likelihood value (white cross) and best-fit value found by the Nelder-Mead optimization algorithm (green cross). (**d**) The likelihood plot shows the best-fit value as well as 95% confidence intervals for the inferred total driver mutation rate from the 'synonymous' variant (neutral mutation) allele frequency spectrum based on the inferred fitness from the 'nonsynonymous' variant allele frequency spectrum.



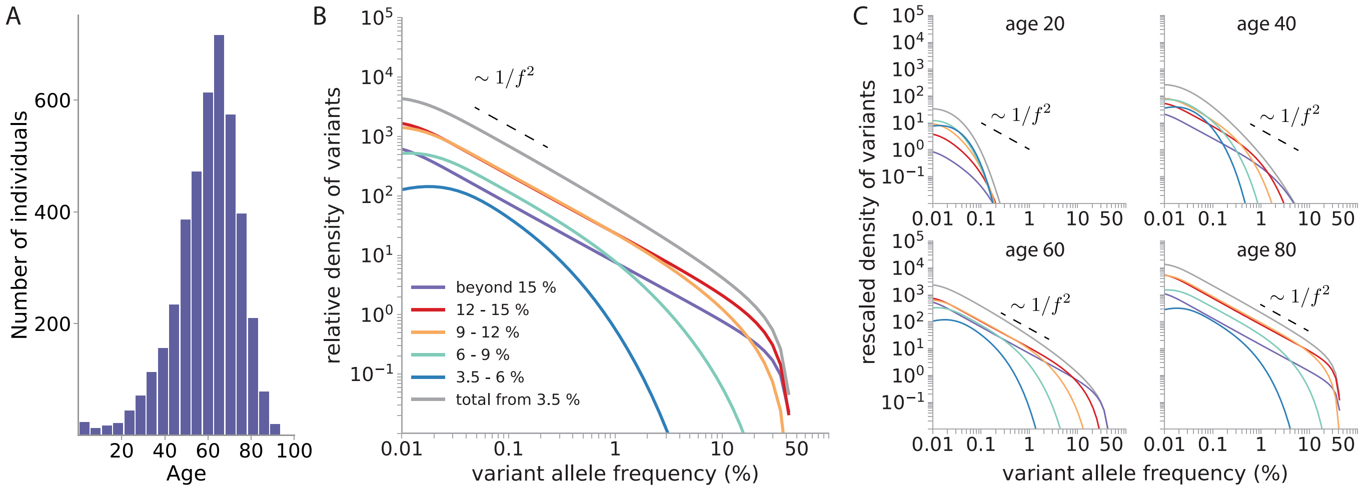
Extended Data Fig. 2 | Developmental mutation rates averages to 2-4 SNVs across entire genome per cell doubling. (**a**) SNV VAFs in HSPC single-cell colonies in an 8 - week foetus[30] where coverage is 22.6x per colony. SNVs found between 35% - 65% (within the dashed lines) are likely clonal in the colony. (**b**) SNV VAFs in HSPC single-cell colonies in an 18 - week foetus[30] where coverage is 12.2x per colony. SNVs found between 30% - 70% (within the dashed lines) are likely clonal in the colony. (**c**) The best-fit to the reverse cumulative for the number of mutations per cell doubling per haploid is 1.86 (95% CI = 1.6 - 2.1) for Lee Six et al. data (green line and datapoints), 1.0 (95% CI = 1.0-1.1) for Chapman et al. 8-week foetus (purple line and datapoints) and 1.0 (95% CI = 0.9-1.1) for Chapman et al. 18-week foetus (orange line and datapoints).



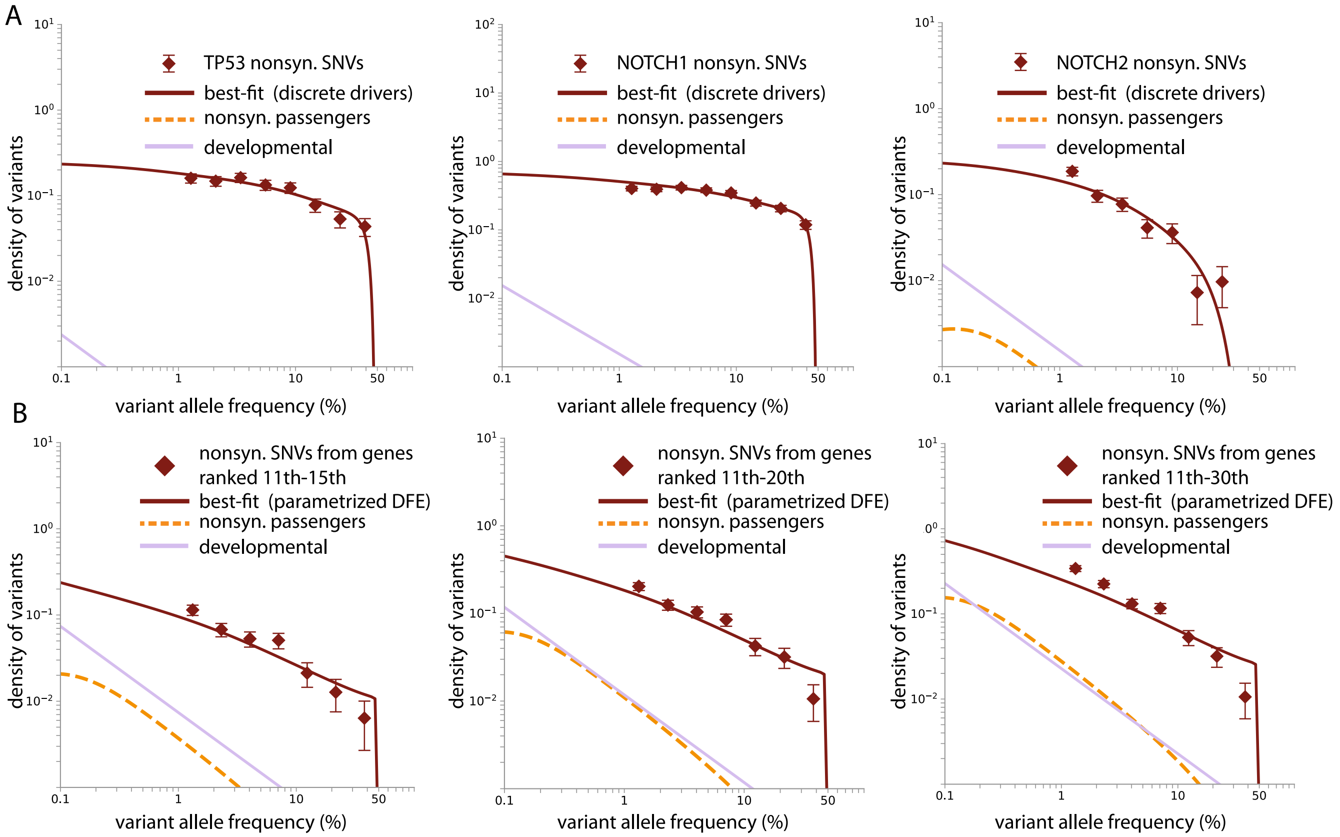
Extended Data Fig. 3 | Inferring the unobserved driver mutation rate using nonsynonymous VAF spectra in Bolton et al. The best-fit nonsynonymous VAF spectrum based on the distribution of ages in the cohort (n = 4160) includes nonsynonymous developmental contribution estimated by considering sizes of the genomic regions (light purple line, Supplementary note 3B) and possible nonsynonymous passengers (orange dashed lines). (**a**) Best-fit haploid driver rate of the most commonly mutated gene (DNMT3A) is 2.9×10^{-6} per year based on the DFE defined by equation 18 (Supplementary note 3C). (**b**) Best-fit haploid driver rate of the top 5 genes (DNMT3A, TET2, PPM1D, SF3B1, ATM) is 4.1×10^{-6} per year. (**c**) Best-fit haploid driver rate of the top 10 genes (DNMT3A, TET2, PPM1D, SF3B1, ATM, ASXL1, JAK2, TP53, SRSF2, CHEK2) is 4.8×10^{-6} per year. Data are presented as mean values ± sampling error.



Extended Data Fig. 4 | Mutation rates of missing drivers assuming different fitness effects. (**a**) The higher the fitness effects of the unobserved drivers, the lower the mutation rate needed to explain the discrepancy in the synonymous VAF density. Inset: Pie chart showing the fraction of explained, unexplained positive selection by observed drivers (all nonsynonymous SNVs on the panel[15]) and developmental contribution to the observed synonymous VAF spectra. (**b**) The observed synonymous VAF spectra (data points, variant number = 344) compared to the density predicted by observed drivers and developmental mutations (dashed orange line) and the predicted density by also including unobserved drivers with different fitness effects (solid orange lines). Data are presented as mean values ± sampling error.



Extended Data Fig. 5 | Contribution from different parts of the DFE to the predicted passenger spectrum. (**a**) The age distribution of the 4160 individuals in Bolton et al.[15]. (**b**) The predicted passenger spectrum in healthy blood according to the inferred distribution of fitness effects in healthy blood (Supplementary note 3C, 'p = 3') and best-fit total driver mutation rate from the synonymous VAF spectrum in blood (Supplementary note 3E) for the age distribution of the 4160 individuals. (**c**) The relative contribution to the passenger spectrum of driver mutations with different fitness effects changes as the individual ages. The total (grey line) represents the passenger VAF spectrum contributed by all driver mutations whose fitness s > 3.5%, below which contribution to the passenger spectrum is very small.



Extended Data Fig. 6 | Nonsynonymous VAF spectra in Martincorena et al. (**a**) The nonsynonymous VAF spectra of the top 10 genes (ranked by nonsynonymous SNV occurrence) were analyzed based on N $\tau = 7800$ (Supplementary note 4C) to estimate their respective fitness and mutation rates. The analysis treats the distribution of fitness effects as delta functions each with a single-valued mutation rate and fitness, taking into account developmental contribution and possible passengers among nonsynonymous SNVs. (**b**) The nonsynonymous VAF spectra of genes beyond the top 10 (ranked by nonsynonymous SNV occurrence) were analyzed based on the chosen DFE (Supplementary note 3C). Similarly, developmental contribution and possible passengers among nonsynonymous SNVs were taken into account. Data are presented as mean values ± sampling error.

nature research

Corresponding author(s): Gladys Poon, Dr Caroline Watson, Professor Daniel Fisher, Dr Jamie Blundell

Last updated by author(s): Jun 14, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Microsoft Excel
Data analysis	Python version 3.7.1 and R version 3.6.1 codes used for data analysis can be found on Github: https://github.com/blundelllab/Genetic-hitchhiking. Nelder-Mead algorithm was used for optimizing fit.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The principal dataset Bolton et al. can be downloaded using the link: https://raw.githubusercontent.com/papaemmelab/bolton_NG_CH/master/M_long.txt. Razavi et al. can be downloaded from the European Genome-Phenome Archive (EGA) archive under accession no. EGAS00001003755. All synonymous variants analyzed in this manuscript are listed in Supplementary tables 1, 2 and 3. The sequencing data for healthy oesophagus was originally reported in Martincorena et al. and may be found in the EGA under accession codes EGAD00001004158 and EGAD00001004159 and can be downloaded directly from here: https://www.science.org/doi/suppl/10.1126/science.aau3879/suppl_file/aau3879_tables2.xlsx

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study is based on data of other publicly available studies and involves only computational and mathematical analyses.
Research sample	Please refer to Bolton et al. and Martincorena et al.
Sampling strategy	Please refer to Bolton et al. and Martincorena et al.
Data collection	Please refer to Bolton et al. and Martincorena et al.
Timing and spatial scale	Please refer to Bolton et al. and Martincorena et al.
Data exclusions	Individuals without hematologic malignancies are included for the main blood analyses. No data was excluded from the Martincorena dataset.
Reproducibility	Please refer to Bolton et al. and Martincorena et al.
Randomization	Please refer to Bolton et al. and Martincorena et al.
Blinding	Please refer to Bolton et al. and Martincorena et al.

Did the study involve field work? □ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging