Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects

Neil A. Robertson[1#], Eric Latorre-Crespo[1#], Maria Terradas-Terradas[2,3], Alison C. Purcell[2,3], Benjamin J Livesey[1], Joseph A. Marsh[1], Lee Murphy[4], Angie Fawkes[4], Louise MacGillivray[4], Mhairi Copland[2], Riccardo E. Marioni[5], Sarah E. Harris[6,7], Simon R. Cox[6,7], Ian J. Deary[6,7], Linus J. Schumacher[8*], Kristina Kirschner[2,3*], Tamir Chandra[1*]

1. MRC Human Genetics Unit, University of Edinburgh, Edinburgh, EH4 2XU, UK
2. Institute of Cancer Sciences, University of Glasgow, Glasgow, G61 1BD, UK
3. Cancer Research UK Beatson Institute, Glasgow, UK
4. Edinburgh Clinical Research Facility, University of Edinburgh, Edinburgh, EH4 2XU, UK
5. Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK
6. Lothian Birth Cohorts, Department of Psychology, The University of Edinburgh, Edinburgh, UK
7. Department of Psychology, The University of Edinburgh, Edinburgh, UK
8. Centre for Regenerative Medicine, University of Edinburgh, Edinburgh, EH16 4UU, UK

# Equal Contribution
*Correspondence to: linus.schumacher@ed.ac.uk, kristina.kirschner@glasgow.ac.uk, tamir.chandra@igmm.ed.ac.uk

Abstract

The prevalence of clonal haematopoiesis of indeterminate potential (CHIP) in healthy individuals increases rapidly from age 60 onwards and has been associated with increased risk for malignancy, heart disease and ischemic stroke. CHIP is driven by somatic mutations in stem cells that are also drivers of myeloid malignancies. Since mutations in stem cells often drive leukaemia, we hypothesised that stem cell fitness substantially contributes to transformation from CHIP to leukaemia. Stem cell fitness is defined as the proliferative advantage over cells carrying no or only neutral mutations. It is currently unknown whether mutations in different CHIP genes lead to distinct fitness advantages that could form the basis for patient stratification. We set out to quantify the fitness effects of CHIP drivers over a 12 year timespan in older age, using longitudinal error-corrected sequencing data. We developed a new method based on drift-induced fluctuation (DIF) filtering to extract fitness effects from longitudinal data, and thus quantify the growth potential of variants within each individual. Our approach discriminates naturally drifting populations of cells and faster growing clones, while taking into account individual mutational context. We show that gene-specific fitness differences can outweigh inter-individual variation and therefore could form the basis for personalised clinical management.

Introduction

Age is the single largest factor underlying the onset of many cancers (de Magalhaes, 2013). Age-related accumulation and clonal expansion of cancer-associated somatic mutations in healthy tissues has been posited recently as a pre-malignant status consistent with the multistage model of carcinogenesis (Martincorena, 2019). However, the widespread presence of cancer-associated mutations in healthy tissues highlights the complexity of early detection and diagnosis of cancer (Ayachi et al., 2020; Genovese et al., 2014; Jaiswal et al., 2014; Lee-Six et al., 2019; Martincorena et al., 2015).

Clonal haematopoiesis of indeterminate potential (CHIP) is defined as the clonal expansion of haematopoietic stem and progenitor cells (HSPCs) in healthy aged individuals. CHIP affects more than 10% of individuals over the age of 60 years and is associated with an estimated 10-fold increased risk for the later onset of haematological neoplasms (Ayachi et al., 2020; Genovese et al., 2014; Jaiswal et al., 2014). There is a clear benefit of detecting CHIP early as the association between clone size and malignancy progression is well-established (Jaiswal and Ebert, 2019; Jaiswal et al., 2014; Park and Bejar, 2020).

The particular mechanisms by which common mutations of CHIP, e.g. *DNMT3A*, *TET2*, contribute to the progression of leukaemia are still not understood, which hinders early diagnosis of CHIP on a gene or variant-basis (Challen and Goodell, 2020; Jaiswal and Ebert, 2019; Shih et al., 2012; Terradas-Terradas et al., 2020). In clinical practice, CHIP is diagnosed by the presence of somatic mutations at variant allele frequencies (VAF) of at least 0.02 in cancer-associated genes in more than 4% of all blood cells (Jaiswal and Ebert, 2019; Steensma and Bolton, 2020). Clonal fitness, defined as the proliferative advantage of stem cells carrying a mutation over cells carrying no or only neutral mutations, has emerged as an alternative clone-specific quantitative marker of CHIP (Watson et al., 2020; Williams et al., 2020). As mutations in stem cells often drive leukaemia (Jaiswal et al., 2014), we hypothesise that stem cell fitness contributes substantially to transformation from CHIP to leukaemia.

Stratification of individuals to inform clinical management for early detection or prevention of leukaemia will depend on our ability to accurately associate genes and their variants with progression to disease. However, it remains unresolved whether variant- or gene-specific fitness effects transcend other factors contributing to variable progression between individuals such as environment or genetics.

Hitherto fitness effects have been predicted from large cross-sectional cohort data (Abelson et al., 2018; Watson et al., 2020). In this approach, single time-point data from many individuals is pooled to generate allele frequency distributions. Although this method allows the study of a large collection of variants, pooling prevents estimation of an individual's mutational fitness effects from cross-sectional data. Inferring fitness from a single time-point creates additional uncertainty about whether a mutation has arisen recently and has grown rapidly (high fitness advantage), or arose a long time ago and grown slowly (low fitness advantage). With longitudinal samples, fitness effects of individual mutations can be estimated directly from the change in VAF over multiple time-points.

In this study we work with longitudinal data from the Lothian Birth Cohorts of 1921 (LBC1921) and 1936 (LBC1936). Such longitudinal data are rare worldwide owing to their participants' older age (70-90 years) and their three-yearly follow-ups over 12 years in each cohort and over 21 years total timespan. We developed a new framework for extracting fitness effects from longitudinal data. Firstly, a drift-induced fluctuations (DIF) filter allowed us to better segregate between naturally drifting populations of cells and fast-growing clones. Secondly, a context-dependent fitting method quantified the growth potential or fitness effects simultaneously for all selected mutations within each individual. We detected gene-specific fitness effects within our cohorts, highlighting the potential for personalised clinical management.

Material and Methods

Participant samples

The Lothian Birth Cohort 1921 (LBC1921) contains a total of 550 participants at Wave 1 of their testing (done between 1999 and 2001) with a gender ratio of 234/316 (m/f) and a mean age at Wave 1 of 79.1 (SD=0.6) (Table 1, (Taylor et al., 2018)). The Lothian Birth Cohort 1936 (LBC1936), contains a total of 1091 participants at Wave1 of their testing (done between 2004 and 2007) with a gender ratio of 548/543 (m/f) and a mean age at Wave 1 of 69.5 (SD=0.8) (Table 1, (Taylor et al., 2018)). We previously identified 73 participants with CHIP at Wave 1

(Robertson et al., 2019). We sequenced DNA from those 73 LBC participants longitudinally and added 16 LBC participants with previously unidentified CHIP, for a total of 298 samples together with 14 “Genome in a Bottle” (GIAB) controls (Supplemental Table 1).

Targeted, error-corrected Sequencing and Data filtering

DNA was extracted from EDTA whole blood using the Nucleon BACC3 kit, following the manufacturer’s instructions. Libraries were prepared from 200ng of each DNA sample using the Archer VariantPlex 75 Myeloid gene panel and VariantPlex Somatic Protocol for Illumina Sequencing (Invitae, Table 2), including modifications for detecting low allele frequencies. Sequencing of each pool was performed using the NextSeq 500/550 High-Output v2.5 (300 cycle) kit on the NextSeq 550 platform. To inform reproducibility, background model for error and batch correction, we sequenced “genome in a bottle” DNA in each batch of samples (DNA NA12878 Coriell Institute).

Reads were filtered for phred ≥30 and adapters removed using Trimmomatic (Bolger et al., 2014) before undergoing guided alignment to human genome assembly hg19 using bwa-mem and bowtie2. Unique molecular barcodes (ligated prior to PCR amplification) were utilised for read deduplication to support quantitative multiplexed analysis and confident mutation detection. Within targeted regions, variants were called using three tools (Lofreq (Wilm et al., 2012), Freebayes (Garrison and Marth, 2012) and Vision (ArcherDX, unpublished) – building a consensus from the output of all callers (Table 3).

All filtered variants at 0.02 VAF met the following criteria: 1) the number of reads supporting the alternative allele surpasses the coverage criteria while exhibiting no directional biases (AO ≥ 5, UAO ≥ 3); 2) variants are significantly underrepresented within the Genome Aggregation Database (gnomAD; $p \leq 0.05$); 3) variants are not obviously germline variants (stable VAF across all waves ~0.5 or ~1) that may have been underrepresented in the gnomAD database due to the narrow geographical origin of the Lothian Birth Cohort participants; 4) contain events that are overrepresented across the dataset - generally frameshift duplications and deletions - whose reads share some sequence homology to target regions yet are likely misaligned artefact from the capture method. In addition, we manually curated this list checking for variants that were previously reported, as per Jaiswal et al. (Jaiswal et al., 2014b), in the COSMIC (Catalogue of Somatic Mutations in Cancer) database or within the published literature (Table 4). Finally, for any variant that surpassed the above criteria at, or above, 0.02 VAF across the measured time period, we endeavored to include all other participant matched data points regardless of VAF level (SuppFig.1A).

Computational prediction of missense variant effects

To predict which missense variants are most likely to be damaging, we used six computational variant effect predictors recently identified as being most useful for identifying pathogenic mutations (Carter et al., 2013; Hecht et al., 2015; Ioannidis et al., 2016; Livesey and Marsh, 2020; Raimondi et al., 2017; Riesselman et al., 2018; Vaser et al., 2016). Specifically, for each variant identified in this study, we determined what fraction of previously identified pathogenic and likely pathogenic missense variants from ClinVar, and what fraction of variants observed in the human population from gnomAD v2.1 for each computational predictor. We then averaged these fractions across all predictors. Note that DeepSequence was only included for DNMT3A and TP53 due to its computational intensiveness and difficulty of running on long protein sequences.

Modelling to estimate fitness

Given the longitudinal nature of this study we can directly fit the deterministic solution of an established minimal model of cell division (Till et al., 1964; Watson et al., 2020) to the time evolution of VAF trajectories in a participant's genetic profiles (Fig.2C). For each individual we simultaneously estimate the fitness and time of acquisition of variants as well as the size of the stem cell pool.

In this model cells exist in two states: stem cells (SCs) or differentiated cells (DCs). Under the assumption that DCs cannot revert to a SC state, differentiation inevitably leads to cell death and is treated as such. Furthermore, assuming that each SC produces the same amount of fully differentiated blood cells allows a direct comparison between the VAF of a variant as observed in blood samples and the number of SCs forming the genetic clone (clone size). For an individual with a collection of clones $\{c_i\}_{i\in I}$, the VAF evolution in time $v_i(t)$ of a clone c_i corresponds to $v_i(t) = \frac{n_i(t)}{2N(t)}$, where $v_i(t)$ is the variant allele frequency of the variant at time t, $n_i(t)$ is the number of SCs carrying the variant and $N(t)$ corresponds to the total number of diploid HSPCs present in the individual. Finally, we assume that that $N(t) = N_w + \sum_{i\in I} n_i(t)$ where N_w is the average number of wildtype (WT) HSPCs in the individual.

The bias towards self-renewal of symmetric divisions is parameterised by s and determines the fitness advantage of a clone. In normal haematopoiesis $s = 0$, in which case clones undergo neutral drift. The stochastic time-evolution of neutral clones corresponds to that of a critical birth and death model (CBM) that we exploit to identify mutations with a clear proliferative advantage, as detailed below. For clones with non-neutral (fitness-increasing) mutations, $s > 0$, and these clones grow in size exponentially in time as $e^{s(t-t_0)}$ from an initial population of 1 SC at the time of mutation acquisition t_0.

Altogether, the exponential proliferation of an isolated genetic clone in an individual results in the logistic time-evolution of its corresponding VAF,

$$v(t) = \frac{1}{2 + 2N_w e^{-s(t-t_0)}},$$

allowing us to estimate the fitness advantage in trajectories with at least 3 data points - 46% of all fit trajectories - using least-squares fitting. Further, when multiple fit clones are present in an individual we constrain the fit to share the stem cell pool size $N(t)$ for all variant trajectories in this individual. This increases the data/parameter ratio, and produces richer dynamics, where exponentially growing clones can be suppressed by the growth of a fitter clone. Critically, this implies that even non-competitive models, where trajectories grow independently of each other, will result in competitive dynamics in the observed VAF trajectories as variants strive for dominance of the total production of blood cells.

Notice that this model cannot account for loss of heterozygosity events. Overall, only one variant - *JAK2 c.1849G>T* - averaged a VAF > 0.5 and was consequently left out of this study. Further, following the fit, three filters were implemented to detect inaccurately fitted trajectories left out from the reported fitness estimates: 1) individuals presenting a large r^2-value, total squared error accumulated over all fitted trajectories in a participant (2 of 54 participants); 2) trajectories with a negative Pearson correlation coefficient between fit and observed data, e.g. a growing trajectory with a decreasing fit, which could result from fitter variants that were not captured in our gene panel (34 of 90 trajectories when using a 0.02 VAF filter, 8 of 96 trajectories when using a drift-induced fluctuations filter, see below and Results for filtering approaches); 3) individuals with clinical records indicating treatment that could affect the clonal dynamics (1 participant, see SI methods). Both exclusion criteria 1 and 2 included participants that likely had clones with multiple mutations based on visual inspection.

Filtering fit variants using mathematical modelling of neutral drift

A critical birth-death process (Fig.2C) predicts drift-induced fluctuations (DIF) in the size of clones with zero mean and variance $\sigma(t)^2 = 2\lambda n(t_0)t$, where $n(t_0)$ is the clone size at the start of the observation period of length t, and λ is the rate of symmetric self-renewing cell divisions. This results in a predicted linear relationship between the VAF of a clone, v, and the expected variance in the distribution of its fluctuations, Δv (VAF gradient between two time points normalised to be independent of time, SI methods). More precisely, the distribution of DIF, Δv, will have expectation $\mu_{\Delta v} = 0$ and variance $\sigma_{\Delta v}{}^2 = v(t_0)\lambda/N$, where N is again the total number of HSPCs in the individual. In the presence of a sufficiently large growing non-neutral mutation, the expected size of DIF of neutral clones is no longer zero but decreases with VAF (SI methods), $\mu_{\Delta v} \sim -v(t_0)$, while $\sigma_{\Delta v}{}^2$ remains proportional to VAF (SI methods).

We use the longitudinal trajectories of synonymous mutations, as a proxy for neutral mutations, in the LBC to fit linear regressions $\mu_{\Delta v} \sim -v(t_0)$ and $\sigma_{\Delta v}{}^2 \sim v(t_0)$ (Fig.3B). The DIF filter then identifies variants whose growth cannot be explained by neutral drift fluctuations, that is variants v satisfying $\Delta v > \mu_{\Delta v} + 2\sigma_{\Delta v}$(SI methods).

Results

The Lothian Birth Cohorts allow for longitudinal profiling of CHIP variants in advanced age

The Lothian Birth Cohorts (LBCs) of 1921 (n=550) and 1936 (n=1091) are two independent, longitudinal studies of ageing with approximately three yearly follow up for five waves, from the age of 70 (LBC1936) and 79 (LBC1921) (Taylor et al., 2018). We previously identified 73 participants with CHIP at Wave 1 through whole-genome sequencing (Robertson et al., 2019). Here, we used a targeted error-corrected sequencing approach using a 75 gene panel (ArcherDX/Invitae) to assess longitudinal changes in variant allele frequencies (VAF) and clonal evolution over 21 years across both LBC cohorts and 12 years within each cohort (Table 1). Error-corrected sequencing allowed accurate quantification, providing more sensitive clonal outgrowth estimates compared to our previous WGS data. We sequenced 298 samples (89 individuals across 2-5 time-points) and achieved a sequencing depth of 1953x mean coverage (1930x median) over all targeted sites with an average of 2.4 unique somatic variants (pan-cohort VAF 0.0005-0.87, median VAF 0.025) detected per participant. We examined all participant-matched events across the time-course: sequence quality control metrics revealed

Figure 1.

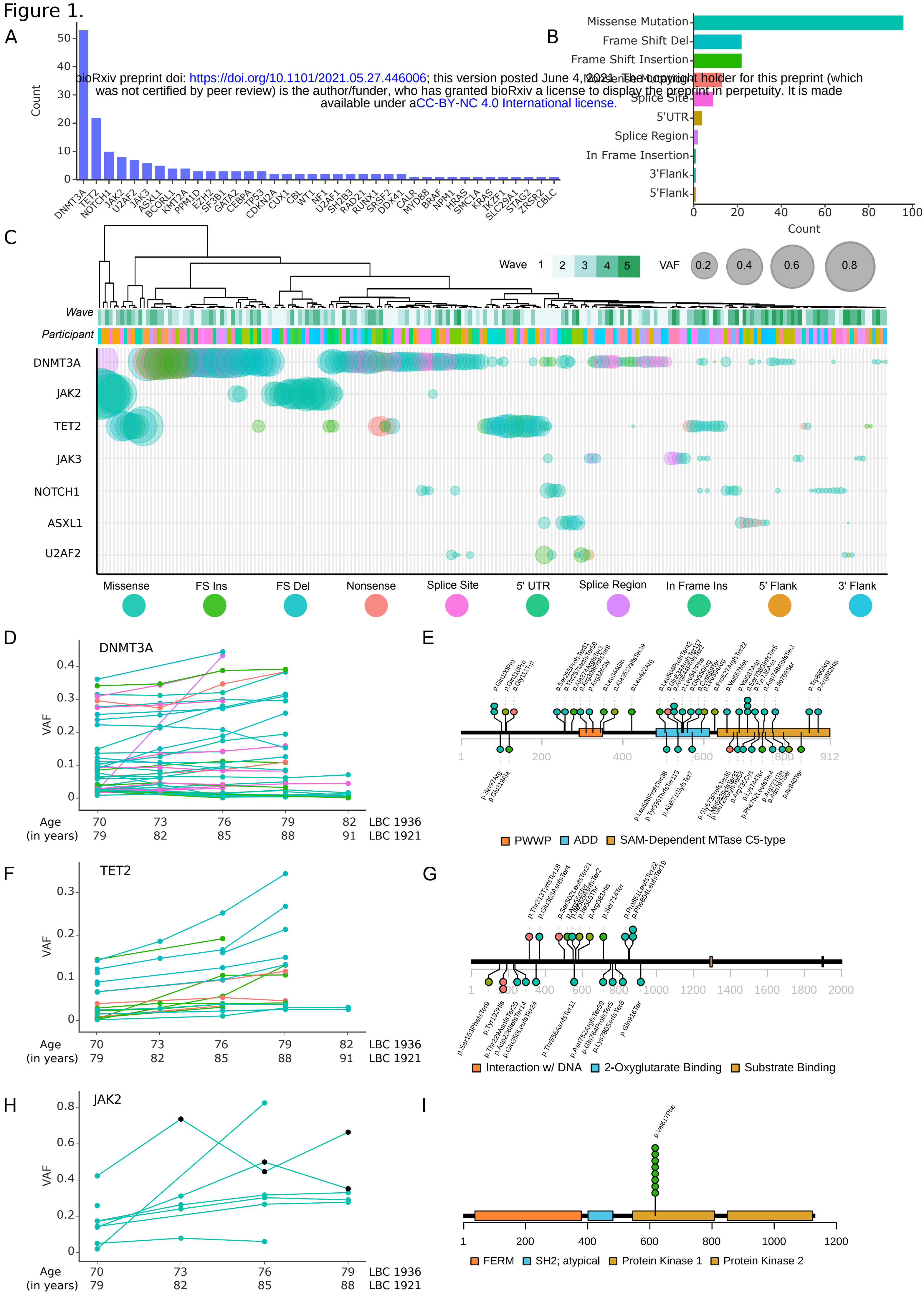

A
Count
0
10
20
30
40
50
DNMT3A
TET2
NOTCH1
JAK2
U2AF2
JAK3
ASXL1
BCORL1
KMT2A
PPM1D
EZH2
SF3B1
GATA2
CEBPA
TP53
CDKN2A
CUX1
CBL
WT1
NF1
U2AF1
SH2B3
RAD21
RUNX1
SRSF2
DDX41
CALR
MYD88
BRAF
NPM1
HRAS
SMC1A
KRAS
IKZF1
SLC29A1
STAG2
ZRSR2
CBLC
B
Missense Mutation
Frame Shift Del
Frame Shift Insertion
Nonsense Mutation
Splice Site
5'UTR
Splice Region
In Frame Insertion
3'Flank
5'Flank
0
20
40
60
80
100
Count
C
Wave
1
2
3
4
5
VAF
0.2
0.4
0.6
0.8
Wave
Participant
DNMT3A
JAK2
TET2
JAK3
NOTCH1
ASXL1
U2AF2
Missense
FS Ins
FS Del
Nonsense
Splice Site
5' UTR
Splice Region
In Frame Ins
5' Flank
3' Flank
D
DNMT3A
VAF
0
0.1
0.2
0.3
0.4
Age
(in years)
70
79
73
82
76
85
79
88
82
91
LBC 1936
LBC 1921
E
p.Gln100Pro
p.Gln110Pro
p.Gly113Trp
p.Ser255ProfsTer61
p.Thr257MetfsTer59
p.Ala274ArgfsTer3
p.Arg309ProfsTer8
p.Arg326Gly
p.Leu344Gln
p.Ala353ValfsTer39
p.Leu422Arg
p.Leu504ProfsTer42
p.Gln534ArgfsTer117
p.Arg544ProfsTer2
p.Leu547Phe
p.Gly550Arg
p.Cys559Ter
p.Leu594Arg
p.Pro627ArgfsTer22
p.Val657Met
p.Val687Asp
p.Ser708GlnfsTer5
p.Tyr735Asn
p.Asp748AlafsTer3
p.Ile769Ser
p.Trp860Arg
p.Arg882His
1
200
400
600
800
912
p.Ser97Arg
p.Glu119Ala
p.Leu508ProfsTer38
p.Tyr536ThrfsTer115
p.Ala571GlyfsTer7
p.Gly673ProfsTer35
p.Met682IlefsTer31
p.Glu725GlyfsTer54
p.Arg736Cys
p.Lys744Ter
p.Phe752LeufsTer4
p.Arg771Gln
p.Asn797Ser
p.Ile840Ter
PWWP
ADD
SAM-Dependent MTase C5-type
F
TET2
VAF
0
0.1
0.2
0.3
Age
(in years)
70
79
73
82
76
85
79
88
82
91
LBC 1936
LBC 1921
G
p.Thr313TyrfsTer18
p.Glu368AsnfsTer4
p.Ser502LeufsTer31
p.Arg550Ter
p.Ile565AsnfsTer2
p.Ile565Thr
p.Arg581His
p.Ser714Ter
p.Pro851LeufsTer22
p.Phe854LeufsTer19
1
200
400
600
800
1000
1200
1400
1600
1800
2000
p.Ser153PhefsTer9
p.Tyr192His
p.Thr229AsnfsTer25
p.Asp236IlefsTer14
p.Glu350LeufsTer24
p.Thr556AsnfsTer11
p.Asn752ArgfsTer59
p.Gln764ProfsTer5
p.Lys780SerfsTer8
p.Gln916Ter
Interaction w/ DNA
2-Oxyglutarate Binding
Substrate Binding
H
JAK2
VAF
0
0.2
0.4
0.6
0.8
Age
(in years)
70
79
73
82
76
85
79
88
LBC 1936
LBC 1921
I
p.Val617Phe
1
200
400
600
800
1000
1200
FERM
SH2; atypical
Protein Kinase 1
Protein Kinase 2

Figure 1. Clonal Haematopoiesis in the Lothian Birth Cohorts of 1921 and 1936.

A. Counts of unique events that exceeded 0.02 VAF across the range of the longitudinal cohorts (approx. 70 - 79 years and 79 - 91 years in LBC36 and LBC21, respectively) in our panel of 75 haematopoietic genes.

B. Counts of the functional consequences of the unique events listed in Fig.1A, highlighting missense mutations as the most frequently encountered event type. Several other key protein altering event types rank highly, including frameshift insertions/deletions and nonsense mutations.

C. Schematic of the top seven most affected genes in the cohort with the largest clone size of an event in any given gene shown. All affected participants have been clustered across all time-points, with the point size scaled by VAF and coloured by the functional consequence of the variant (as per Fig.1B and legend). Participants broadly cluster together across their time-course, driven by the expanding or stable VAF of their harboured mutations and underscores the high prevalence and large clone size of common clonal haematopoietic drivers, namely, DNMT3A, TET2 and JAK2.

D. Clone size trajectories of all DNMT3A mutations across the time-series in both LBC21 and LBC36 coloured by the functional consequence of the variant (as per Fig.1B and 1C).

E. Locations of the somatic mutations discovered in DNMT3A. Protein affecting events are marked and labeled across the structure of the gene (missense in red, truncating in purple, stacked for multiple events) highlighting a strong, previously described tendency for CH events in DNMT3A to occur in conserved structural and domain loci (which are coloured and labeled along the amino acid length of the protein).

F. Clone size trajectories of all TET2 mutations across the time-series in both LBC21 and LBC36 coloured by the functional consequence of the variant (as per Fig.1B and 1C).

G. The locations of somatic mutations in TET2. Protein affecting events are marked and labeled across the structure of the protein (missense in red, truncating in purple, stacked for multiple events).

H. Clone size trajectories of all JAK2 mutations across the time-series in both LBC21 and LBC36 coloured by the functional consequence of the variant (as per Fig.1B and 1C). Points marked in black denote time-points after which the affected participant received treatment for leukaemia - potentially driving the observed reductions in clone size.

I. The locations of somatic mutations in JAK2. Protein affecting events are marked and labeled across the structure of the protein (missense in red, truncating in purple, stacked for multiple events). Here, all eight JAK2 mutations are p.Val617Phe (JAK2-V617F) missense variants.

that only 12 of 466 data-points failed to meet our quality criteria likely due to low initial VAF. The majority of our variant loci generally displayed a high number of supporting reads, with a mean of 186 (SuppFig.1A).

For our initial analysis, we retained variants with at least one time-point at 0.02 VAF (Table 5). *DNMT3A* was the most commonly mutated CHIP gene (n=53 events in 42 participants), followed by *TET2* (n=22 events in 18 participants), *NOTCH1* (n=10 events in 9 participants), *JAK2* (n=8 events in 8 participants) and *U2AF2* (n= 7 events in 6 participants) (Fig.1A and 1C). Our mutation spectrum is consistent with previous studies in finding *DNMT3A* and *TET2* as the most frequently mutated genes. However, we detected a lower frequency of mutations in splicing genes, such as *SRSF2, U2AF1* and *SF3B1*, despite the older age of the cohort (Fig.1A and SuppFig.1A). This is in contrast to previously published cohort data, where splicing mutations became more prominent with increased age (McKerrell et al., 2015). The majority of mutations were missense, frameshift and nonsense mutations (Fig.1B). Clonal evolution and changes in VAF per participant over all time points are shown in Fig.1C and SuppFig.1C. We detected some variants more frequently at certain hot spots within a gene such as p.Arg882His in *DNMT3A*, with previously unreported variants being present as well (Fig.1D-I, Table 2, (Jaiswal et al., 2014)). In the case of *JAK2V617F*, we identified two individuals who developed leukaemia at Wave 2 and received treatment between Waves 2 and 3, likely driving a clear reduction in clone size (Fig.1H). Those indivdiuals were excluded from further analysis. Overall, our sequencing approach allowed for high resolution, longitudinal mapping of CHIP variants over a 12-year time span per cohort and 21 year timespan across both cohorts from the same geographical region and born 9 years apart.

Identification of harmful CHIP variants in the LBCs

Stem cell fitness is defined as the proliferative advantage over cells carrying no or only neutral mutations. It remains incompletely understood to what extent fitness is gene- or variant-specific, or determined by the bone marrow microenvironment and clonal composition. Earlier estimates suggested a wide spread of fitness effects even for variants of the same gene (Watson et al., 2020), which would make it difficult to clinically stratify individuals with CHIP. To determine the fitness effects of the variants identified in our cohorts (Fig.1A), we initially selected all CHIP variants in our data using the commonly used criterion of defining any variants with VAF>0.02 as CHIP (Jaiswal and Ebert, 2019; Steensma and Bolton, 2020), and retaining only those variants with at least 2 time-points (Fig.2A). This approach identified 140 CHIP mutations

Figure 2.

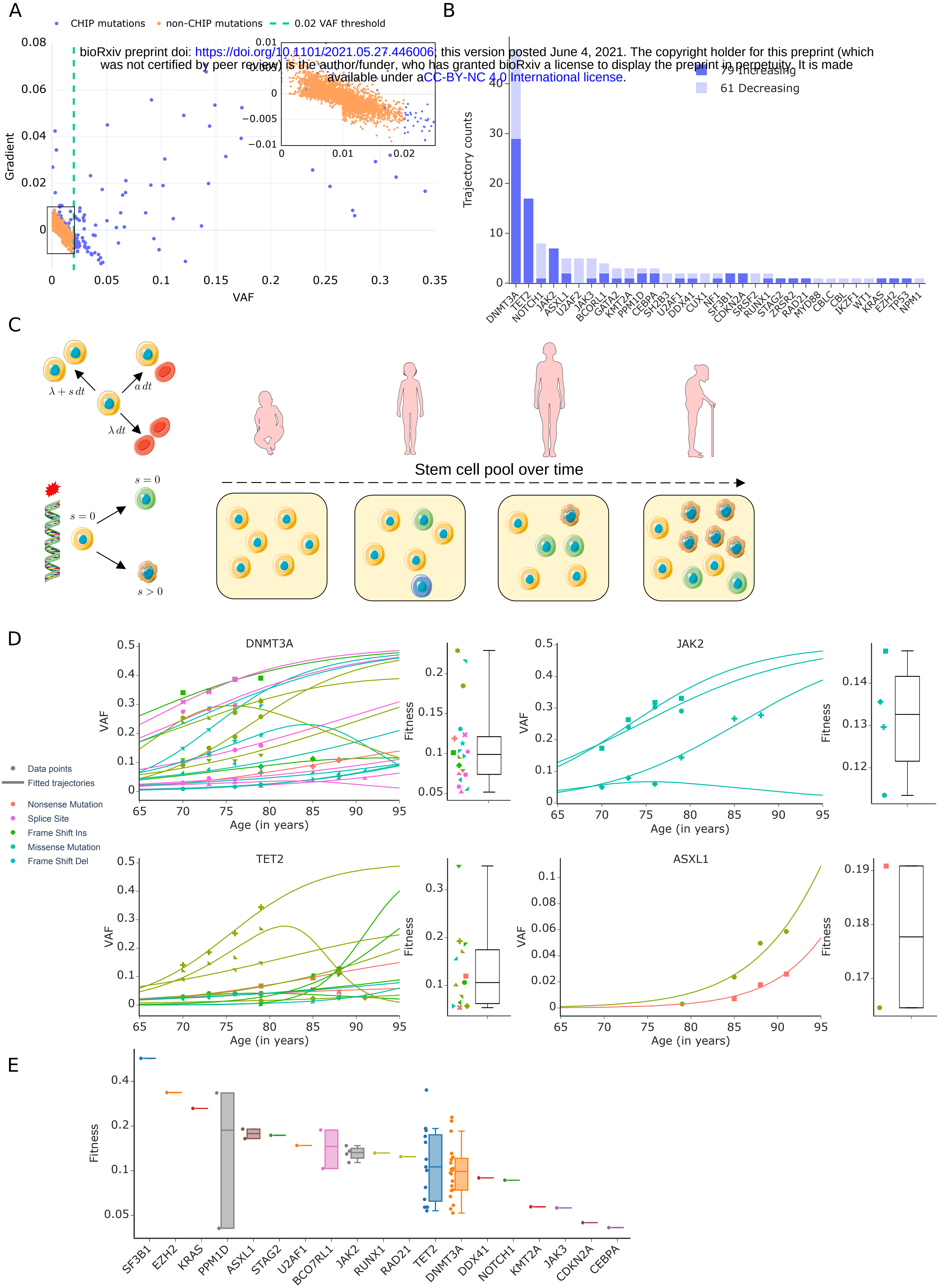

A
CHIP mutations
non-CHIP mutations
0.02 VAF threshold
Gradient
VAF
0.08
0.06
0.04
0.02
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.01
0.005
0
−0.005
−0.01
0
0.01
0.02
B
79 Increasing
61 Decreasing
Trajectory counts
40
30
20
10
0
DNMT3A
TET2
NOTCH1
JAK2
ASXL1
U2AF2
JAK3
BCORL1
GATA2
KMT2A
PPM1D
CEBPA
SH2B3
U2AF1
DDX41
CUX1
NF1
SF3B1
CDKN2A
SRSF2
RUNX1
STAG2
ZRSR2
RAD21
MYD88
CBLC
CBL
IKZF1
WT1
KRAS
EZH2
TP53
NPM1
C
$\lambda + s\,dt$
$a\,dt$
$\lambda\,dt$
$s = 0$
$s = 0$
$s > 0$
Stem cell pool over time
D
Data points
Fitted trajectories
Nonsense Mutation
Splice Site
Frame Shift Ins
Missense Mutation
Frame Shift Del
DNMT3A
JAK2
TET2
ASXL1
VAF
Age (in years)
Fitness
65
70
75
80
85
90
95
0.5
0.4
0.3
0.2
0.1
0
0.2
0.15
0.1
0.05
0.14
0.13
0.12
0.3
0.2
0.1
0.1
0.08
0.06
0.04
0.02
0
0.19
0.18
0.17
E
Fitness
0.4
0.2
0.1
0.05
SF3B1
EZH2
KRAS
PPM1D
ASXL1
STAG2
U2AF1
BCO7RL1
JAK2
RUNX1
RAD21
TET2
DNMT3A
DDX41
NOTCH1
KMT2A
JAK3
CDKN2A
CEBPA

Figure 2. Fitness effects of variants thresholded at 0.02 VAF.

A. VAF measurement $v(t_0)$ at initial time-point t_0 vs time-normalized gradient in VAF, $\left(v(t_{end}) - v(t_0) \right) / \sqrt{t_{end} - t_0}$, between initial and last time-points t_0 and t_{end} of all variants detected in the LBC with at least 2 time-points. Each data point corresponds to a trajectory in the LBC and has been coloured according to its CHIP status based on the 0.02 VAF threshold (dashed green line). Blue and orange respectively denote whether or not trajectories achieved a VAF > 0.02 during the observed time-span. Inset: Close-up of low VAF region showing a large number of variants below 0.02 VAF. Note data below 0.01 VAF were only included if the trajectory later rises above 0.01, hence the data are less dense below VAF=0.01.

B. Number of trajectories passing the currently used 0.02 VAF-threshold, broken down into whether VAF is increasing or decreasing from first to last time-point.

C. Schematic of mathematical model of clonal dynamics. Stem cells naturally acquire mutations over time leading to the formation of genetic clones. Artwork includes images by Servier Medical Art licensed under CC BY 3.0.

D. Fitted trajectories (lines) and fitness estimates (box-plots) for all variants in selected genes DNMT3A, JAK2, ASXL1, TET2. Note decreasing trajectories occur even for non-zero fitness because our model fitting takes into account the presence of other fit mutations in the same individual.

E. Fitness effects of mutations ranked by median fitness. Note fitness is displayed on a logarithmic scale to emphasize relative differences in fitness between variants. Boxplots show median and exclusive interquartile range.

(Fig.2B) overall. To estimate the fitness effect each variant confers, we fitted longitudinal trajectories using birth-death models of clonal dynamics (Fig.2C) on trajectories with 3 or more time-points (Table 6). The resulting fitness values show an overall dependence of fitness on the gene level (Fig.2E), with a wide distribution of fitness for some genes, such as TET2 and DNMT3A (Fig.2D and 2E), but not others such as JAK2 (which are all the same variant).

Longitudinal trajectories allow for a more accurate stratification of CHIP variants

Since longitudinal data allow direct quantification of the growth in VAF over time, we can inspect the gradients (fluctuations) in VAF for variants that were classified as CHIP based on thresholding. We find that a VAF>0.02 threshold not only misses fast growing and potentially harmful variants (Fig.2A), but can include variants whose frequencies are shrinking (Fig.2A and 2B) and thus either do not confer a fitness advantage or are being outcompeted by other clones.

Overall, only 56% of CHIP mutations detected by thresholding at 0.02 VAF were growing during the observed time span (Fig.2B). Longitudinal data thus reveal limitations in defining CHIP mutations based on a widely used VAF threshold. We therefore sought to filter variants based on longitudinal information, by selecting variants whose growth cannot be explained by the drift in VAF from neutral fluctuations (Fig.3A). Using the longitudinal behaviour of synonymous mutations in the LBCs, as a proxy for neutral trajectories, we find that the predicted linear relationships for mean and variance of VAF-gradients (see Methods) accurately reflect the data (Fig.3B). This novel approach, which we named drift-induced fluctuations (DIF) filter, allows us to estimate the distribution of fluctuations due to the neutral drift of clone populations as a function of VAF and detects variants growing more than two standard deviations from the mean of the predicted neutral drift (Fig.3A, Table7). DIF-filtering resulted in 226 variant trajectories (Fig.3C), 97% of which grew over the observed time span. We note that the VAF of fit mutations may still shrink over time due to the presence of an even fitter clone in the same individual. This is in contrast to thresholding at 0.02 VAF, with only 56% of variants identified to be growing and thus likely to confer a fitness advantage. Of 226 variants we detected, only 81 would have been detected using the previous VAF-threshold filter approach. We recomputed fitness estimates for this new set of filtered trajectories (Fig.3F). Growing variants that were missed by the traditional filtering method include highly fit variants such as *U2AF1* c.470A>G and *NPM1* c.847-4C>T (Fig.3D and 3F). VAF-thresholding identified *NOTCH1* as one of the top three most frequently mutated genes in our cohort. However, only 1 of the 8 variants identified was growing over the observed time-course. In contrast, all 6 *NOTCH1* variants identified by using our DIF filter were growing, thus our filtering method allows us to more effectively separate benign from potentially harmful variants. Overall, the variants only detected by DIF-filtering, but not VAF-thresholding, were of high fitness (Fig.3E).

We further stratified variants using a combination of computational predictors (Fig.3G, see Methods), categorising the most prevalent CHIP variants into damaging (6 variants), possibly damaging (3 variants) and likely benign (2 variants), as well as frameshifts and terminations (34 variants, which are also most likely damaging to protein structure and thus protein function). The novel DIF-filter therefore produces a low false discovery rate of pathogenic variants, with 88% of the detected fit variants being predicted to be likely damaging, frameshift or termination. As our approach detects very few benign variants by design, we were unable to report statistically significant differences between the fitness distributions classified by damaging. Further,

Figure 3.

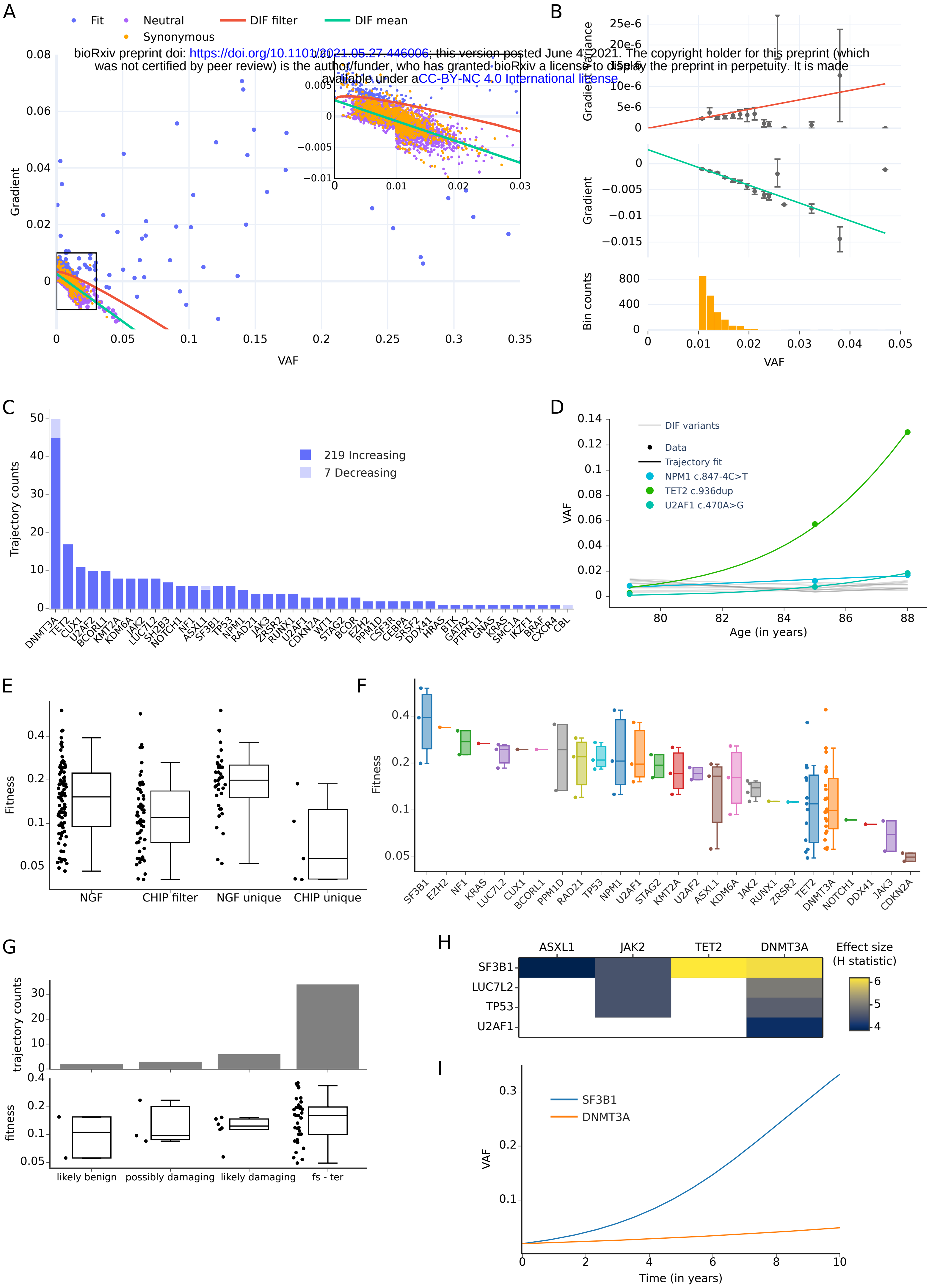

A
Fit
Neutral
Synonymous
DIF filter
DIF mean
Gradient
VAF
0.08
0.06
0.04
0.02
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.005
0
−0.005
−0.01
0.01
0.02
0.03
B
Gradient variance
25e-6
20e-6
15e-6
10e-6
5e-6
Gradient
−0.005
−0.01
−0.015
Bin counts
800
400
0.04
0.05
C
Trajectory counts
50
40
30
20
10
219 Increasing
7 Decreasing
DNMT3A
TET2
CUX1
U2AF2
BCORL1
KMT2A
KDM6A
JAK2
LUC7L2
SH2B3
NOTCH1
NF1
ASXL1
SF3B1
TP53
NPM1
RAD21
JAK3
ZRSR2
RUNX1
U2AF1
CDKN2A
WT1
STAG2
BCOR
EZH2
PPM1D
CSF3R
CEBPA
SRSF2
DDX41
HRAS
BTK
GATA2
PTPN11
GNAS
KRAS
SMC1A
IKZF1
BRAF
CXCR4
CBL
D
DIF variants
Data
Trajectory fit
NPM1 c.847-4C>T
TET2 c.936dup
U2AF1 c.470A>G
0.14
0.12
0.1
0.08
0.06
Age (in years)
80
82
84
86
88
E
Fitness
0.4
0.2
0.1
0.05
NGF
CHIP filter
NGF unique
CHIP unique
F
Fitness
SF3B1
EZH2
NF1
KRAS
LUC7L2
CUX1
BCORL1
PPM1D
RAD21
TP53
NPM1
U2AF1
STAG2
KMT2A
U2AF2
ASXL1
KDM6A
JAK2
RUNX1
ZRSR2
TET2
DNMT3A
NOTCH1
DDX41
JAK3
CDKN2A
G
trajectory counts
30
20
10
fitness
likely benign
possibly damaging
likely damaging
fs - ter
H
ASXL1
JAK2
TET2
DNMT3A
SF3B1
LUC7L2
TP53
U2AF1
Effect size
(H statistic)
6
5
4
I
SF3B1
DNMT3A
VAF
0.3
0.2
0.1
Time (in years)
0
2
4
6
8
10

Figure 3. Filtering trajectories based on whether they exceed neutral fluctuations.

A. Time-normalised gradient in VAF vs VAF for variants detected in the LBC with at least 2 time-points and at least one VAF >0.01 per trajectory, with synonymous (orange dots), non-synonymous fit (blue dots) and likely neutral (purple dots) mutations, classified based on whether the gradient exceeds the drift-induced fluctuations (DIF) of synonymous mutations (red line, two standard deviations above mean, green line). Inset: Close-up of low VAF region. Note that data below 0.01 VAF were only included if the trajectory later rises above 0.01, hence the data are less dense below VAF=0.01.

B. Statistics of fluctuations of synonymous mutations. Both the mean (middle) and variance (top) of time-normalised VAF-gradients were fitted with a linear model. Only data at VAF ≥ 0.01 where used as data below this are incompletely sampled (see A). The fit was weighted by the number of data in each bin (bottom).

C. Number of trajectories passing the DIF filter, broken down into whether VAF is increasing or decreasing from first to last time-point.

D. Variant trajectories in one individual classified according to the DIF filter. Mutations with positive fitness effects are coloured according to variant classification (see Fig.1B) while likely neutral trajectories are shown in grey. For fit variants, points correspond to the observed VAF measurements and lines to the model fit. While the TET2 variant was also detected using the 0.02 VAF-threshold, the NPM1 and U2AF1 were only detected using the DIF-filter, as they did not reach 0.02 VAF. In particular the U2AF1 variant is fast growing (s=0.364).

E. Fitness effects of variants broken down by filtering method.

F. Fitness effects of mutations ranked by median fitness. Note fitness is displayed on a logarithmic scale to emphasize relative differences in fitness between variants (E&F). Boxplots show median and exclusive interquartile range

G. Fitness effects of variants broken down by predicted mutation effect (see Methods).

H. Analysis of variance of the distribution of fitness across genes. Heatmap of all statistically significant (p<0.05) Kruskal-Wallis H statistics, labeled by effect size, computed for all combinations of pairs of genes.

I. Predicted growth of SF3B1 and DNMT3A variants over 10 years, all starting at 0.02 VAF, using the median fitness effects per gene (F).

predicted benign mutations selected as fit could be the result of clonal hitchhiking (mutations co-occurring in cells carrying a damaging mutation).

Taken together, measuring changes in synonymous variants combined with mathematical models of neutral drift results in an alternative method, DIF-filtering, that improves on the threshold-based definition of CHIP mutations (Fig.3A), by replacing an arbitrary cut-off on VAF by a choice of false discovery rate and as a result selecting fewer trajectories with shrinking VAF (Fig.2B and 3C). Importantly, only 2 time points are necessary to apply DIF-filtering, making this a widely applicable method for existing cohorts and future studies.

Predicting trajectories using our fitness estimates from single-time point data

We further analysed differences in the distributions of fitness between genes using a non-parametric test. Despite having small sample sizes for many genes we still detected statistically significant differences among the distributions of fitness effects (Fig.3H). In particular, we found that mutations in gene *SF3B1* conferred a higher fitness advantage over mutations in common genes such as *DNMT3A, TET2* or *ASXL1*. Differences in the distribution of fitness allow us to predict the future growth of mutations from initial time-points. For example, if a patient presents with a variant in gene *SF3B1* or *DNMT3A* at 0.02 VAF, their growth is predicted to differ by 158% percent in 3 years (respectively achieving 0.057 and 0.026 VAF, assuming median fitness for each gene), warranting a clinical follow-up over that time-frame. We have also tested differences in fitness by genes when summarised into functional categories and found trajectories of genes involved in DNA methylation to have lower fitness than genes involved in splicing, the cohesin complex and histone methylation (SuppFig.2G and 2H).

Discussion

The clinical potential for stratifying progression of CHIP depends on whether genes confer distinct fitness advantages. Indeed, most studies so far have not shown a clear distinction of fitness effects on a gene basis, and have shown considerable overlap in fitness coefficients between variants of different genes. We show that fitness can substantially differ by gene and gene category. Combining longitudinal data with a new method to identify CHIP variants allows for more accurate fitness estimates of CHIP than cross sectional cohort data, and motivates further studies with increased sample sizes.

The strength of our approach, combining longitudinal data with our new DIF-filtering method, is exemplified by *NOTCH1* and *LUC7L2*. *NOTCH1* variants identified by a 0.02 VAF threshold were prominent in the LBCs (Fig.1A). However, all but one of these were shrinking in frequency. In contrast, our DIF-filtering method identified 6 variants conferring a fitness advantage, including the only increasing variant detected at a 0.02 VAF level, c.4598A>C (Fig.3C). Although *NOTCH1* mutations commonly occur in leukaemia, this specific mutation has not been reported in COSMIC, thus more cohort studies are needed to clarify a role for *NOTCH1* in CHIP (Ferrando, 2009; Rossi et al., 2012). *LUC7L2* is a pre-mRNA splicing factor that has been previously reported to be involved in CHIP, MDS and AML (Hershberger et al., 2016; Sperling et al., 2017). *LUC7L2* was excluded from our list of CHIP mutations when using a 0.02 VAF cut-off. In contrast, we find several *LUC7L2* mutations using DIF-filtering, thus supporting previous findings. Other well-described CHIP genes that were only detected using DIF-filtering include *NPM1, TP53* and *U2AF2* (see Fig.2E and 3F,(Jaiswal et al., 2014)). DIF-filtering also results in higher fitness estimates for some genes, such as *PPM1D, U2AF1* and *RAD21*, because we either detected more trajectories of a gene with the DIF filter or were able to detect other variants at small VAF in the same individuals (Fig.2E and 3F). Importantly, a minimum of two time points per participant is needed to apply DIF-filtering, making this method widely applicable to future prospective studies and existing cohorts.

Synonymous mutations in our study reached a VAF of up to 0.02 (Fig.3B). Some previous studies have claimed that neutral events will typically not expand to > 0.001 VAF over a human lifespan (Lee-Six et al., 2018), and that all synonymous variants detected are likely hitchhikers (Watson et al., 2020). Our longitudinal data clearly show many synonymous and non-synonymous mutations larger than 0.001 VAF that do not grow in VAF (Fig.3A) or whose growth and shrinkage over time is consistent with neutral drift (Fig.3B and SI methods). If the synonymous mutations were able to grow to such high frequencies by hitchhiking on clones with fit mutations, these no longer have a fitness advantage over the observation period. An alternative explanation considers the number N of self-renewing cells (HSPCs). For an individual aged *t*, neutrally drifting mutations are only expected to achieve $\sim \sqrt{t}/N$ VAF. Recent studies, therefore, suggest that the high frequency at which CHIP is found in the elderly is likely the result of a small pool of stem cells (Zink et al., 2017). Estimates for the number of HSPCs in healthy adult humans vary widely, with recent phylogenetic estimates suggesting 50k-200k (Lee-Six et al., 2018), whereas other estimates suggest 11-22k (Abkowitz et al., 2002). On the lower end, a study based on allometric scaling estimates that the “pool of actively replicating cells that contribute to hematopoiesis” is around 500 cells (Dingli and Pacheco, 2006). In our study, the distribution of gradients of synonymous variants provides the estimate $\theta \sim \lambda/N$, where θ corresponds to the slope of the linear relation between variance and vaf (Fig.3B) and λ to the rate of self-renewing divisions. Using the well-established rate $\lambda = 1.3$/year (once every

40 weeks, (Catlin et al., 2011)) we obtain an estimate on the number of HSPCs of $N \sim 6500$ (see SI methods for a more detailed analysis). The presence of neutral events reaching large VAFs in our data could be explained by a depletion of the stem cell pool in participants of the LBC (80 years mean across all collected blood samples). Both a study of somatic mutations in a 115-yr-old woman and a recent longitudinal study following the evolution of clones in an individual from age 103 to 110 (3 time points) found that likely benign clones achieved VAFs of up to 0.3 (van den Akker et al., 2020; Holstege et al., 2014)- suggesting a very small pool of stem cells. Another possible explanation is that the observed variants arose in cells with faster rates of self-renewing symmetric divisions λ, either due to natural selection of faster dividing HSPCs or because the observed variants in the LBC arose in different HSPC compartments, such as multipotent progenitor cells (Barile et al., 2020; Morcos et al., 2020; Takahashi et al., 2021). Since we can only directly measure θ, it is possible that both a depletion of the pool of HSPCs at old age and heterogeneity in the self-renewal rate λ have enabled neutral events to reach high VAFs. Despite the disparity of estimates on the number of HSPCs in healthy individuals between studies, the fitness estimates we report (Fig.2 and Fig.3) are valid for the wide ranges of reported HSPC estimates (SuppFig.2E and 2F and SI methods).

In this work, we have assumed each mutation to uniquely identify a clone. It is possible for multiple mutations to have occurred in the same clone, though this has been estimated to be as rare as 0.16 double mutants per person, for mutations conveying a fitness advantage $s = 0.1$, achieving VAF > 0.005 at age 80 (Watson et al., 2020). There may be a few variants that are double mutants, and these would increase the distribution of fitness reported for individual genes (Fig.2C and 3F), but not strongly affect the median fitness per gene. Double mutant clones in which one mutation is a hitchhiker contributing little to the fitness effect could be identified by seeking trajectories with very similar fitness values. Further work could estimate the fitness effects of double mutants by identifying possible subclone structures and then fitting models separately for each case. Ultimately, we will need single-cell targeted sequencing or colony growth essays combined with sequencing to verify the identity of clones with multiple mutations.

Acknowledgments and author contributions

We thank Prof. Chris P. Ponting for critical reading of the manuscript.
The LBC1921 was supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC), a Royal Society–Wolfson Research Merit Award to I.J.D., and the Chief

Scientist Office (CSO) of the Scottish Government's Health Directorates. The LBC1936 is supported by Age UK (Disconnected Mind project, which supports S.E.H), the Medical Research Council (MR/M01311/1, MR/K026992/1, which supported I.J.D.), and the University of Edinburgh. KK is funded by a John Goldman Fellowship sponsored by Leukaemia U.K. (2019/JGF/003). M.T.T. and N.R. are supported by MRC funded Ph.D. studentships (MR/N013166/1). TC and LS are supported by Chancellor's Fellowships held at the University of Edinburgh. JAM is a Lister Institute Research Fellow. ELC is a cross-disciplinary post-doctoral fellow supported by funding from the University of Edinburgh and Medical Research Council (MC_UU_00009/2). We are grateful for funding from the Howat Foundation. LS, KK and TC conceived and supervised the study. NAR, ELC, LS, KK and TC wrote the manuscript. LM, AF, LMG generated data. NAR and ELC developed the methodology for data analysis. NAR, ELC, MTT, ACP, JAM, BJL conducted data analysis. SEH, SC, IJD curated the LBCs and gave access to samples and data. MC advised on aspects of the study.

Conflict of interest

KK received a reagent grant from ArcherDX/ Invitae. LM consults for Illumina.

Data Availability

Data are deposited in the European Genome-Phenome Archive and provided via the LBC DAC (https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration)

Supplemental Tables

Supplemental tables are available at the following link:
https://drive.google.com/file/d/10UMVtRk8rKosBJE3fjnOMpbYyrvgjD_K/view?usp=sharing

References

Abelson, S., Collord, G., Ng, S.W.K., Weissbrod, O., Mendelson Cohen, N., Niemeyer, E., Barda, N., Zuzarte, P.C., Heisler, L., Sundaravadanam, Y., et al. (2018). Prediction of acute myeloid leukaemia risk in healthy individuals. Nature *559*, 400–404.

Abkowitz, J.L., Catlin, S.N., McCallie, M.T., and Guttorp, P. (2002). Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. Blood *100*, 2665–2667.

van den Akker, E.B., Makrodimitris, S., Hulsman, M., Brugman, M.H., Nikolic, T., Bradley, T., Waisfisz, Q., Baas, F., Jakobs, M.E., de Jong, D., et al. (2020). Dynamic clonal hematopoiesis and functional T-cell immunity in a supercentenarian. Leukemia.

Ayachi, S., Buscarlet, M., and Busque, L. (2020). 60 Years of clonal hematopoiesis research: From X-chromosome inactivation studies to the identification of driver mutations. Exp. Hematol. *83*, 2–11.

Barile, M., Busch, K., Fanti, A.-K., Greco, A., Wang, X., Oguro, H., Zhang, Q., Morrison, S.J., Rodewald, H.-R., and Höfer, T. (2020). Hematopoietic stem cells self-renew symmetrically or gradually proceed to differentiation. BioRxiv 2020.08.06.239186.

Bailey N.T.J. (1990). The elements of stochastic processes with applications to the natural sciences. J. Wiley and Sons v25.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics *14*, S3.

Catlin, S.N., Busque, L., Gale, R.E., Guttorp, P., and Abkowitz, J.L. (2011). The replication rate of human hematopoietic stem cells in vivo. Blood *117*, 4460–4466.

Challen, G.A., and Goodell, M.A. (2020). Clonal hematopoiesis: mechanisms driving dominance of stem cell clones. Blood *136*, 1590–1598.

Dingli, D., and Pacheco, J.M. (2006). Allometric Scaling of the Active Hematopoietic Stem Cell Pool across Mammals. PLoS One *1*, e2.

Ferrando, A.A. (2009). The role of NOTCH1 signaling in T-ALL. Hematol. Am. Soc. Hematol. Educ. Progr. 353–361.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. ArXiv Prepr. ArXiv1207.3907.

Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. N. Engl. J. Med. *371*, 2477–2487.

Hecht, M., Bromberg, Y., and Rost, B. (2015). Better prediction of functional effects for sequence variants. BMC Genomics *16*, S1.

Hershberger, C.E., Hosono, N., Singh, J., Dietrich, R.C., Gu, X., Makishima, H., Sauntharajah, Y., Maciejewski, J.P., and Padgett, R.A. (2016). The Role of LUC7L2 in Splicing and MDS. Blood *128*, 5504.

Holstege, H., Pfeiffer, W., Sie, D., Hulsman, M., Nicholas, T.J., Lee, C.C., Ross, T., Lin, J., Miller, M.A., Ylstra, B., et al. (2014). Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. Genome Res. *24*.

Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885.

Jaiswal, S., and Ebert, B.L. (2019). Clonal hematopoiesis in human aging and disease. Science (80-.). *366*, eaan4673.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P. V, Mar, B.G., Coleman Lindsley, R., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes A BS TR AC T. NEJM.Org. N Engl J Med *26*, 2488–2498.

Lee-Six, H., Øbro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K., Belmonte, M., Osborne, R.J., Huntly, B.J.P., Martincorena, I., Anderson, E., et al. (2018). Population dynamics of normal human blood inferred from somatic mutations. Nature *561*.

Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., et al. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. Nature *574*.

Livesey, B.J., and Marsh, J.A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol. Syst. Biol. *16*, e9380.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., et al. (2015). High burden and pervasive positive selection of somatic mutations in normal human skin. Science (80-.). *348*, 880 LP – 886.

McKerrell, T., Park, N., Moreno, T., Grove, C.S., Ponstingl, H., Stephens, J., Understanding Society Scientific Group, C., Crawley, C., Craig, J., Scott, M.A., et al. (2015). Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. Cell Rep. *10*, 1239–1245.

Morcos, M.N.F., Li, C., Munz, C.M., Greco, A., Dressel, N., Reinhardt, S., Dahl, A., Becker, N.B., Roers, A., Höfer, T., et al. (2020). Hematopoietic lineages diverge within the stem cell compartment. BioRxiv 2020.08.21.261552.

Park, S.J., and Bejar, R. (2020). Clonal hematopoiesis in cancer. Exp. Hematol. *83*, 105–112.

Raimondi, D., Tanyalcin, I., Ferté, J., Gazzo, A., Orlando, G., Lenaerts, T., Rooman, M., and Vranken, W. (2017). DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. *45*, W201–W206.

Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat. Methods *15*, 816–822.

Robertson, N.A., Hillary, R.F., McCartney, D.L., Terradas-Terradas, M., Higham, J., Sproul, D., Deary, I.J., Kirschner, K., Marioni, R.E., and Chandra, T. (2019). Age-related clonal haemopoiesis is associated with increased epigenetic age. Curr. Biol. *29*.

Rossi, D., Rasi, S., Fabbri, G., Spina, V., Fangazio, M., Forconi, F., Marasca, R., Laurenti, L., Bruscaggin, A., Cerri, M., et al. (2012). Mutations of NOTCH1 are an independent predictor of survival in chronic lymphocytic leukemia. Blood *119*, 521–529.

Shih, A.H., Abdel-Wahab, O., Patel, J.P., and Levine, R.L. (2012). The role of mutations in epigenetic regulators in myeloid malignancies. Nat. Rev. Cancer *12*, 599–612.

Sperling, A.S., Gibson, C.J., and Ebert, B.L. (2017). The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. Nat. Rev. Cancer *17*, 5–19.

Steensma, D.P., and Bolton, K.L. (2020). What To Tell Your Patient With Clonal Hematopoiesis And Why: Insights From Two Specialized Clinics. Blood.

Takahashi, M., Barile, M., Chapple, R.H., Tseng, Y., Nakada, D., Busch, K., Fanti, A.-K., Säwén, P., Bryder, D., Höfer, T., et al. (2021). Reconciling Flux Experiments for Quantitative Modeling of Normal and Malignant Hematopoietic Stem/Progenitor Dynamics. Stem Cell Reports *16*, 741–753.

Taylor, A.M., Pattie, A., and Deary, I.J. (2018). Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. Int. J. Epidemiol. *47*, 1042-1042r.

Terradas-Terradas, M., Robertson, N.A., Chandra, T., and Kirschner, K. (2020). Clonality in haematopoietic stem cell ageing. Mech. Ageing Dev. *189*, 111279.

Till, J.E., McCulloch, E.A., and Simonovich, L. (1964). A Stochastic Model of Stem Cell Proliferation, Based on the Growth of Spleen Colony-Forming Cells. Proc. Natl. Acad. Sci. U. S. A. *51*, 29–36.

Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., and Ng, P.C. (2016). SIFT missense predictions for genomes. Nat. Protoc. *11*, 1–9.

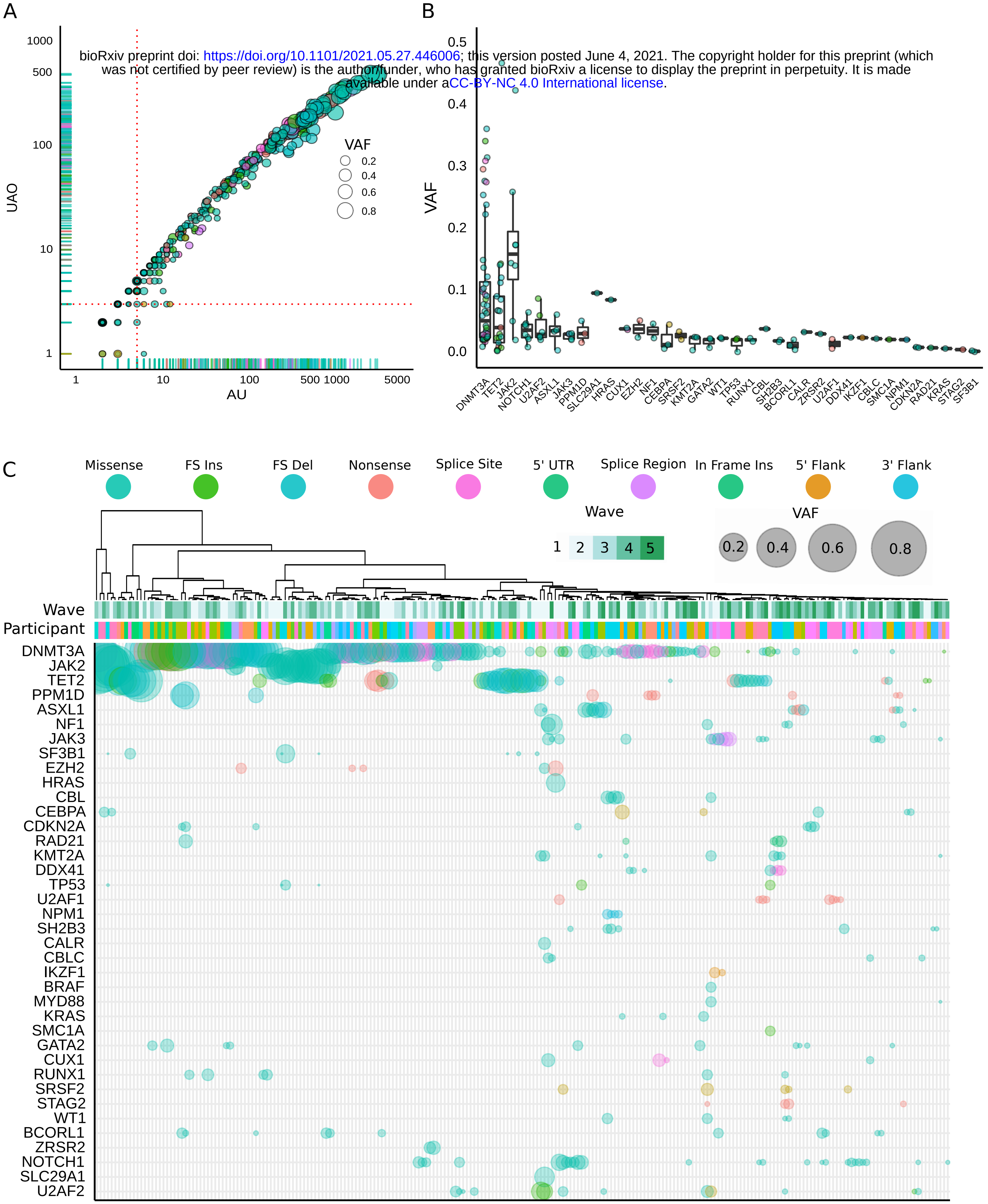
Watson, C.J., Papula, A.L., Poon, G.Y.P., Wong, W.H., Young, A.L., Druley, T.E., Fisher, D.S., and Blundell, J.R. (2020). The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science (80-.). *367*.

Williams, M.J., Zapata, L., Werner, B., Barnes, C.P., Sottoriva, A., and Graham, T.A. (2020). Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dN/dS ratios. Elife *9*, e48714.

Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., and Nagarajan, N. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. *40*, 11189–11201.

Zink, F., Stacey, S.N., Norddahl, G.L., Frigge, M.L., Magnusson, O.T., Jonsdottir, I., Thorgeirsson, T.E., Sigurdsson, A., Gudjonsson, S.A., Gudmundsson, J., et al. (2017). Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. Blood *130*, 742–752.

Figure S1.


A
UAO
AU
1000
500
100
10
1
1
10
100
500 1000
5000
VAF
0.2
0.4
0.6
0.8
B
VAF
0.5
0.4
0.3
0.2
0.1
0.0
DNMT3A
TET2
JAK2
NOTCH1
U2AF2
ASXL1
JAK3
PPM1D
SLC29A1
HRAS
CUX1
EZH2
NF1
CEBPA
SRSF2
KMT2A
GATA2
WT1
TP53
RUNX1
CBL
SH2B3
BCORL1
CALR
ZRSR2
U2AF1
DDX41
IKZF1
CBLC
SMC1A
NPM1
CDKN2A
RAD21
KRAS
STAG2
SF3B1
C
Missense
FS Ins
FS Del
Nonsense
Splice Site
5' UTR
Splice Region
In Frame Ins
5' Flank
3' Flank
Wave
1
2
3
4
5
VAF
0.2
0.4
0.6
0.8
Wave
Participant
DNMT3A
JAK2
TET2
PPM1D
ASXL1
NF1
JAK3
SF3B1
EZH2
HRAS
CBL
CEBPA
CDKN2A
RAD21
KMT2A
DDX41
TP53
U2AF1
NPM1
SH2B3
CALR
CBLC
IKZF1
BRAF
MYD88
KRAS
SMC1A
GATA2
CUX1
RUNX1
SRSF2
STAG2
WT1
BCORL1
ZRSR2
NOTCH1
SLC29A1
U2AF2

Supplemental Figure Legends

Figure S1. Clonal Haematopoiesis in the Lothian Birth Cohorts.

A. Sequence quality metrics for our mutation calls across all participants and time-points filtered for 0.02 VAF. Plotted here are the AO (the number of sequenced reads supporting the alternative allele (mutation)) against the UAO (the number of sequenced reads with unique start sites that support the alternative allele - an additional measure of molecular complexity in our sequencing libraries). The red dotted lines denote the filter thresholds in both measurements (AO ≥ 5, UAO ≥ 3) and points are scaled by the VAF of the somatic mutation. Having mined for participant matched data points for any variant that surpasses our 0.02 VAF threshold, only 12 data points failed to meet our filter criteria (of 466). We did not exclude these as they were supported with matching events across any participants' time series.

B. Box and jitter plot of the variant allele frequency of all observed events in the 1st Wave at a 0.02 VAF filter coloured by variant classification and ordered by largest mean VAF.

C. Schematic of all affected genes in the cohort with the largest clone size of an event in any given gene shown. All affected participants have been clustered across all time-points, with the point size scaled by VAF and coloured by the functional consequence of the variant (as per legend).

Figure S2.

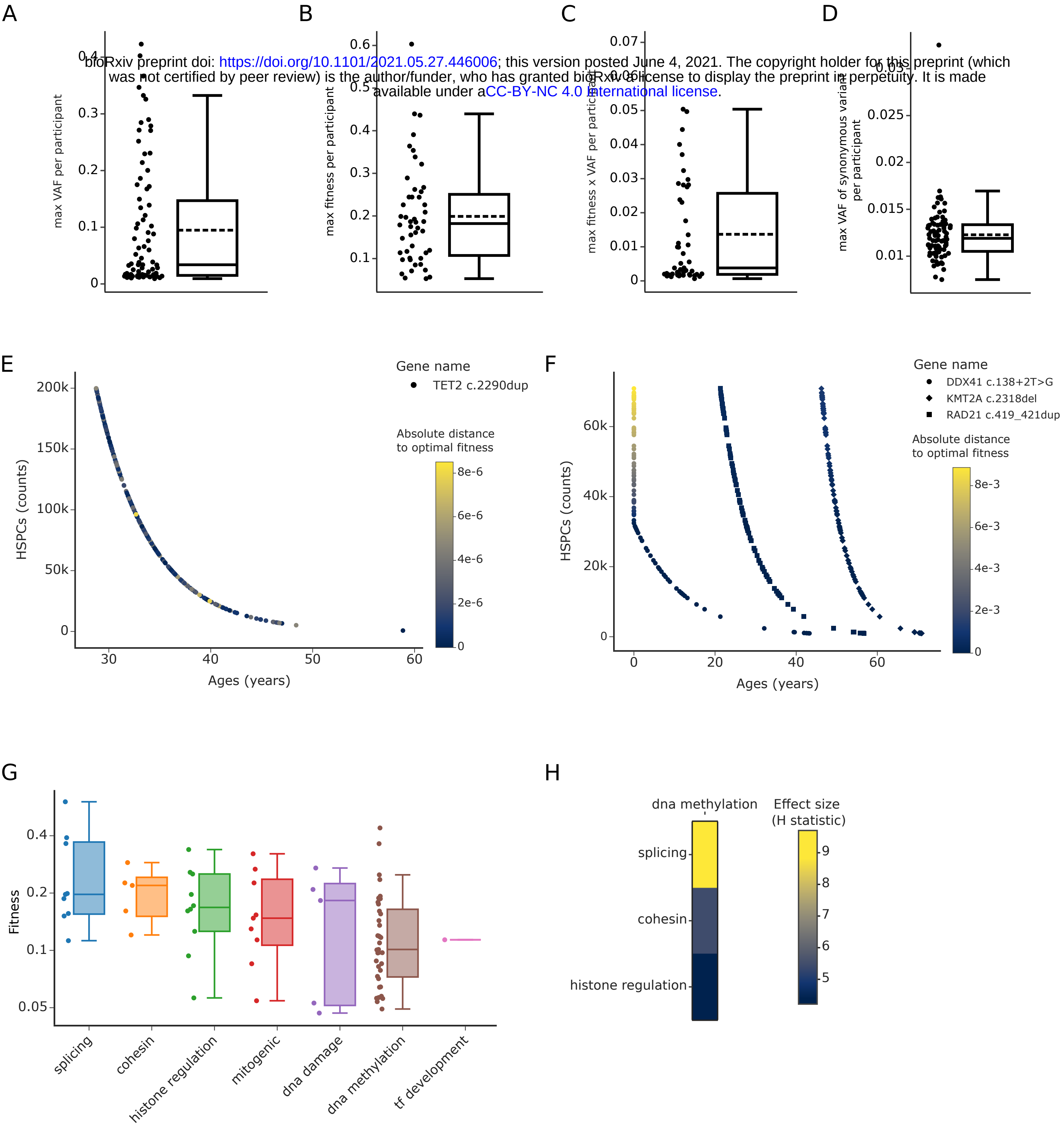

A
max VAF per participant
0.4
0.3
0.2
0.1
0
B
max fitness per participant
0.6
0.5
0.4
0.3
0.2
0.1
C
max fitness x VAF per partic
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
D
max VAF of synonymous varia
per participant
0.03
0.025
0.02
0.015
0.01
E
Gene name
● TET2 c.2290dup
Absolute distance
to optimal fitness
8e-6
6e-6
4e-6
2e-6
0
HSPCs (counts)
200k
150k
100k
50k
0
30
40
50
60
Ages (years)
F
Gene name
● DDX41 c.138+2T>G
◆ KMT2A c.2318del
■ RAD21 c.419_421dup
Absolute distance
to optimal fitness
8e-3
6e-3
4e-3
2e-3
0
HSPCs (counts)
60k
40k
20k
0
0
20
40
60
Ages (years)
G
Fitness
0.4
0.2
0.1
0.05
splicing
cohesin
histone regulation
mitogenic
dna damage
dna methylation
tf development
H
dna methylation
splicing
cohesin
histone regulation
Effect size
(H statistic)
9
8
7
6
5

Figure S2. Cohort properties for fitness estimates

A - D. Box plots show median (bold line) and mean (dashed line) and exclusive interquartile range.

A. Distribution of maximum VAF, v, per participant of the LBC. v is computed as the mean over all observations of a trajectory. One trajectory with $v > 0.5$ was excluded for this distribution.

B. Distribution of maximum inferred fitness, s, per participant of the LBC. This distribution only shows participants with at least one trajectory selected as fit using the DIF filter.

C. Distribution of maximum inferred $s * v$ per participant in the LBC. s corresponds to the inferred fitness of a variant and v to its average observed VAF. This distribution only shows participants with at least one trajectory selected as fit using the DIF filter.

D. Distribution of maximum VAF, v, of synonymous mutations per participant of the LBC. v is computed as the mean over all observations of a trajectory.

E. Hypersurface of parameters producing the same fit to the evolution of a fit clone in a participant of the LBC. Each marker corresponds to a combination of parameters (fitness, number of wild type HSPCs in individual and time of mutation acquisition) fitted to the trajectory of a participant in the LBC. The colour scale shows the absolute difference between the inferred fitnesses associated with each marker and the fit producing the least squared error. All markers shown are within a relative tolerance of 10^{-3} of the solution presenting the least squared error.

F. Hypersurface of parameters producing similar fits to the evolution of clones in a participant of the LBC with 3 fit variants. Each marker corresponds to a combination of parameters (fitness, number of wild type HSPCs in individual and time of mutation acquisition). Curves were fitted simultaneously, sharing a common number of HSPCs. The colour scale shows, for each variant, the absolute difference between the inferred fitnesses associated each marker and to the fit producing the least squared error. All markers shown are within a relative tolerance of 2×10^{-3} of the solution presenting the least squared error. Notice that the fitness of variant *DDX41* marginally increases when the time of mutation acquisition meets the boundary at age 0 in the exponentially decreasing hypersurface of parameter solutions.

G. Fitness effects of mutations grouped by gene category and ranked by median fitness. Note fitness is displayed on a logarithmic scale to emphasize relative differences in fitness between variants.

H. Analysis of variance of the distribution of fitness across genes. Heatmap of all statistically significant (p<0.05) Kruskal-Wallis H statistics, labelled by effect size, computed for all combinations of pairs of genes.

Supplementary Methods for "Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects"

1 Birth-death model of clonal dynamics

Consider a population $X(t)$ of haematopoietic stem and progenitor cells (HSPCs) in an individual at time t. A classical model of stem cell population assumes that HSPCs divide and differentiate according to the following birth-death process, (Till et al., 1964; Watson et al., 2020):

$$\begin{aligned} HSPC &\to HSPC + HSPC, && \text{with} \quad (\lambda + s)\,\mathrm{d}t, \\ &\to D + D, && \text{with} \quad \lambda\,\mathrm{d}t, \\ &\to HSPC + D, && \text{with} \quad A\,\mathrm{d}t, \end{aligned} \tag{1}$$

where D denotes a differentiated cell. Since differentiated cells cannot self-renew, differentiation will result in cell death and is treated as such. The deterministic behaviour of this system is given by

$$\frac{\mathrm{d}X}{\mathrm{d}t}(t) = sX(t). \tag{2}$$

Consequently, the time evolution of the population of HSPCs is

$$X(t) = X(0)e^{st}, \tag{3}$$

where $X(0)$ corresponds to the initial population at time $t = 0$. It is clear that parameter s regulates the excess growth towards self-renewal and dictates the evolution of $X(t)$.

The stochastic behaviour of this model, on the other hand, is more involved. Assuming that $X(0) = 1$, $X(t)$ has the following probability distribution (Bailey, 1990):

$$\mathbb{P}\{X(t) = m\} = \begin{cases} (1-\alpha)(1-\beta)\beta^{m-1}, & m > 0, \\ \alpha, & m = 0, \end{cases} \tag{4}$$

where

$$\alpha = \frac{\lambda \left(e^{st} - 1\right)}{\left(\lambda + s\right) e^{st} - \lambda}, \quad \beta = \frac{\left(\lambda + s\right) \left(e^{st} - 1\right)}{\left(\lambda + s\right) e^{st} - \lambda}. \tag{5}$$

More generally, the distribution accounting for the case when the initial population is $a > 0$,

$$p_m^X\left(t, a\right) = \mathbb{P}\left\{X\left(t\right) = m \mid X\left(0\right) = a\right\}, \tag{6}$$

can be analytically derived as

$$p_m^X\left(t, a\right) = \begin{cases} \sum_{j=0}^{\min(a,m)} \binom{a}{j} \binom{a+m-j-1}{a-1} \alpha^{a-j} \beta^{m-j} \left(1 - \alpha - \beta\right)^j, & m > 0, \\ \alpha^a, & m = 0. \end{cases} \tag{7}$$

Of particular relevance is the mean and variance of $X\left(t\right)$, if $X\left(0\right) = a$:

$$\mu_X\left(t\right) = a e^{st} \quad \text{and} \quad \sigma_X^2\left(t\right) = \frac{a\left(2\lambda + s\right)}{s} e^{st} \left(e^{st} - 1\right). \tag{8}$$

2 A stochastic model of neutral clones

Assume that all HSPCs in an individual follow a critical birth-death (CBD) process ($s = 0$). When a mutation occurs in a HSPC it gives rise to a genetic clone c_i with

$$n_i\left(t\right) = \text{Number of HSPC cells in clone } c_t \text{ at time } t. \tag{9}$$

For now, we assume that mutations have no impact in the bias towards self-renewal, s, and that n_i in turn follows a CBD process. We refer to $\{c_i\}_{i \in \mathcal{C}}$ as the collection of all *neutral clones* ($s = 0$) present in the individual. If we denote by t_i the time of acquisition of the mutation, so that $n_i\left(t_i\right) = 1$, then the probability distribution of $n_i\left(t\right)$ is given by the limiting case $s \to 0$ of (4) , (Bailey, 1990). That is, for any $t > t_i$,

$$\mathbb{P}\left\{n_i\left(t\right) = m\right\} = \begin{cases} \left(\frac{1}{1+\lambda(t-t_i)}\right)^2 \left(\frac{\lambda(t-t_i)}{1+\lambda(t-t_i)}\right)^{m-1}, & m > 0 \\ \frac{\lambda(t-t_i)}{1+\lambda(t-t_i)}, & m = 0, \end{cases} \tag{10}$$

Further, given $t_0 > t_i$, for any $t > t_0$ consider

$$p_m^{n_i}\left(t, a\right) := \mathbb{P}\left\{n_i\left(t + t_0\right) = m \mid n_i\left(t_0\right) = a\right\}. \tag{11}$$

The mean and variance of this probability distribution can be derived, again, as the limit $s \to 0$ of (8),

$$\mu_{n_i}\left(t\right) = n_i\left(t_0\right) \quad \text{and} \quad \sigma_{n_i}^2\left(t\right) = 2\lambda n_i\left(t_0\right)\left(t - t_0\right). \tag{12}$$

2.1 Distribution of VAF gradients in neutral clones

Next, we want to understand how the stochastic evolution of HSPC counts over time translates to the evolution of the variant allele frequency (VAF or blood share) of genetic clones. Further, rather than the distribution of VAFs at a particular time, we want to derive the distribution of expected gradients of VAF between two time points.

2.1.1 Linear scaling of random variables

First, let us remind the scaling properties of the mean and variance of a random variable under affine transformations. Let X be a random variable with mean μ_X and variance σ_X^2 and define Y as

$$Y = \frac{X + b}{k}, \tag{13}$$

for scalars b and k. Then mean and variance scale as follows,

$$\mu_Y = \frac{\mu_X + b}{k} \quad \text{and} \quad \sigma_Y^2 = \frac{\sigma_X^2}{k^2}. \tag{14}$$

2.1.2 Distribution of VAF evolution

Since we assume that all HSPCs in the individual follow a CBD process, the total population of HSPCs remains stable, on average, at N. Further, since asymmetric divisions are far more common than symmetric divisions, we can assume that the number of differentiated blood cells produced by a clone of HSPCs is directly proportional to the clone's size. We therefore model the VAF of a clone c_i at time $t > t_i$ as

$$v_i\left(t\right) = \frac{n_i\left(t\right)}{2N}, \tag{15}$$

where the factor 2 in the denominator is the result of diploidy in HSPCs.

Further, for any time $t > t_0 > t_i$ we can consider the probability distribution $p_m^{v_i}(t, a)$ of the evolution of VAF from an initial state $v_i(t_0) = a$ as in (11). It follows from the distribution of clone sizes (12) and scaling properties (14) that

$$\mu_{v_i}\left(t\right) = v_i\left(t_0\right) \quad \text{and} \quad \sigma_{v_i}^2\left(t\right) = \frac{2\lambda n_i\left(t_0\right)\left(t - t_0\right)}{4N^2}. \tag{16}$$

When observing VAF distributions, it is worth noting the natural emergence of parameter

$$\theta = \frac{\lambda}{N}.$$

A CBD model of the evolution of clones in HSPCs, therefore, yields that the variance of the distribution of VAFs at time t from an initial state $v_i(t_0)$ is

$$\sigma^2_{v_i}(t) = \theta v_i(t_0)\left(t - t_0\right). \tag{17}$$

2.1.3 Distribution of VAF gradients between two time points

Again, for any $t > t_0 > t_i$ consider the VAF evolution $v_i(t)$ of a neutral clone c_i and the resulting probability distribution of the time-adjusted gradient

$$\Delta_{v_i}(t) = \frac{v_i(t) - v_i(t_0)}{\sqrt{t - t_0}}. \tag{18}$$

From scaling properties (14) and the distribution of VAFs of neutral clones (16) and (17), it follows that

$$\mu_{\Delta_{v_i}}(t) = \frac{\mu_{v_i}(t_0) - v_i(t_0)}{\sqrt{t - t_0}} = 0 \tag{19}$$

and

$$\sigma^2_{\Delta_{v_i}}(t) = \frac{\sigma^2_{v_i}(t)}{t - t_0} = \theta v_i(t_0). \tag{20}$$

Notice that by considering a time-adjusted gradient (taking the square root of the time difference) the variance of $\Delta_{v_i}(t)$ becomes independent of time.

3 A stochastic model of neutral clones in the presence of a fit clone

In the following we want to study the evolution of a neutral clone's gradient in the presence of a non-stochastic growing mutation (the deterministic limit for $s \gg 0$). Consider an individual with a collection of neutral clones, $\{c_i\}_{i\in I}$, that harbours a mutation at time t_f giving rise to a clone c_f with a fitness advantage $s > 0$.

In the continuum limit, the deterministic evolution (3) of the clone's size $n_f(t)$ is

$$n_f(t) = e^{s(t-t_f)}.$$

The total pool of HSPCs in this individual is, therefore, comprised of a collection of wild type HSPCs that remains constant in time at N_w and a mutant growing population of size $n_f(t)$:

$$\begin{aligned} N(t) &= \text{expected amount of HSCs at time } t \\ &= N_w + n_f(t). \end{aligned}$$

The evolution of the time-adjusted VAF gradient of neutral clones then becomes

$$\begin{aligned} \Delta_{v_i}(t) &= \frac{v_i(t) - v_i(t_0)}{\sqrt{t-t_0}} \\ &= \frac{1}{2\sqrt{t-t_0}} \left(\frac{n_i(t)}{N(t)} - \frac{n_i(t_0)}{N(t_0)} \right). \end{aligned}$$

3.1 Neutral drift in the presence of a fit clone

Here we will show that $\Delta_{v_i}(t) = a(t)\Delta_{\tilde{v}_i}(t) + b(t)$, where $\tilde{v}_i(t)$ corresponds to the evolution of a neutral clone's VAF in the absence of fit clones,

$$\tilde{v}_i(t) = \frac{n_i(t)}{2N_w}, \tag{21}$$

and $a(t)$ and $b(t)$ are functions of time, independent of the random process $n_i(t)$.

First, notice that

$$\begin{aligned} v_i(t) &= \frac{n_i(t)}{2N(t)} \\ &= f(t)\tilde{v}_i(t), \end{aligned} \tag{22}$$

with $f(t) = N_w/N(t)$. The mean value theorem asserts that $\exists \xi \in [t_0, t]$ so that

$$f(t_0) = f(t) - (t-t_0)f'(\xi) \tag{23}$$

and

$$\begin{aligned} \Delta_{v_i}(t) &= \frac{f(t)\tilde{v}_i(t) - f(t_0)\tilde{v}_i(t_0)}{\sqrt{t-t_0}} \\ &= f(t)\Delta_{\tilde{v}_i}(t) + \sqrt{t-t_0}f'(\xi)\tilde{v}_i(t_0). \end{aligned} \tag{24}$$

3.2 Estimating the mean gradient of neutral variants

Since the expectation operator $\mu_{(\cdot)}(t)$ is a linear operator and $\mu_{\Delta\tilde{v}_i}(t) = 0$, (24) yields

$$\mu_{\Delta_{v_i}}(t) = \sqrt{t-t_0}f'(\xi)\tilde{v}_i(t_0). \tag{25}$$

From the definition of $f(t)$ it follows that, $f'(\xi) = -2sv_f(\xi)f(\xi)$. Combining this expression with (22) we have

$$\mu_{\Delta_{v_i}}(t) = -v_i(t_0)2sv_f(\xi)\frac{f(\xi)}{f(t_0)}\sqrt{t-t_0}, \tag{26}$$

and assuming $f(\xi) \approx f(t_0)$,

$$\mu_{\Delta_{v_i}}(t) \approx -v_i(t_0)2sv_f(\xi)\sqrt{t-t_0} \tag{27}$$

for $\xi \in [t_0, t]$. We use the Lothian Birth cohort (LBC) data to extract statistics about the distributions of VAF and fitness of clones to check that this linear relation holds true (Fig.3B and SuppFig.2A-D). We find that the median over all participants with a measurable fit mutation (fit mutations can occur outside the set of sequenced genes) of the maximum realisation of $v_f(t) * s$ across all present fit clones is

$$\operatorname*{median}_{p \in \text{LBC}} \max_{f \in p} v_f * s \approx 4 \times 10^{-3} \tag{28}$$

where v_f is the averaged VAF over all data points (SuppFig.2C). Further, the average over all combinations of t, t_0 in our data is $\sqrt{t-t_0} \approx 2.5$. With these we estimate

$$\mu_{\tilde{\Delta}v_i}(t) \approx -av_i(t_0), \tag{29}$$

with $a \approx 0.02$.

We estimate the mean of VAF fluctuations of neutral clones by averaging the fluctuations of synonymous trajectories (clonal trajectories of genetic variants not changing the coded amino acid) with the same VAF, v, across all participants, $\mu_{v_i(t_0)=v}(\Delta_{v_i}(t))$. Fitting a weighted linear regression model on our data yields

$$\mu_{v_i(t_0)=v}(\Delta_{v_i}(t)) = -av + b \tag{30}$$

with $a \approx 0.3$ and an intercept $b \approx 10^{-3}$ to account for the lack of observations of extinction events (Fig.3B). There are a number of possible reasons for the numerical discrepancy of the theoretically estimated and empirically observed slope, such as the approximation $f(\xi) \approx f(t_0)$ made above, but also the fact that many individuals will have multiple fit clones affecting the gradients of VAF of neutral variants.

3.3 Estimating the variance of gradients

From (24) and scaling of random variables (14) we find

$$\sigma^2_{\Delta_{v_i}}(t) = f(t)^2\sigma^2_{\Delta_{\tilde{v_i}}}(t). \tag{31}$$

First notice that, in the presence of a small fit clone $n_f(t) < N_w \Leftrightarrow v_f(t) < 0.25$, the scaling factor$f(\xi)$ is bounded by $1/2 < f(\xi)^{-1} < 1$. This initial approximation yields a bound on the variance of fluctuations in the presence of small clones:

$$\frac{\sigma^2_{\Delta_{\tilde{v}_i}}(t)}{4} < \sigma^2_{\Delta_{v_i}}(t) < \sigma^2_{\Delta_{\tilde{v}_i}}(t). \tag{32}$$

A more careful analysis shows a linear relation $\sigma^2_{\Delta_{v_i}}(t) \sim v_i(t_0)$. Combining (31) and (17) we obtain

$$\sigma^2_{\Delta_{v_i}}(t) = \frac{\lambda f(t)^2}{N_w f(t_0)} v_i(t_0). \tag{33}$$

Then, approximating $f(t) \approx f(t_0)$ yields

$$\sigma^2_{\Delta_{v_i}}(t) \approx \frac{\lambda}{N_w + n_f(t)} v_i(t_0). \tag{34}$$

Finally, from the definition of VAF it follows that $N_w + n_f(t) = \frac{N_w}{1-2v_f(t)}$ and

$$\sigma^2_{\Delta_{v_i}}(t) \approx \frac{\lambda(1-2v_f(t))}{N_w} v_i(t_0). \tag{35}$$

To estimate $\sigma^2_{\Delta_{v_i}}(t)$ from the LBC data we approximate $\sigma^2_{\Delta_{v_i}}(t)$ by the variance of observed gradients over all synonymous trajectories with $v_i(t_0) = v$, $\sigma^2_{v_i(t_0)=v}(\Delta_{v_i}(t))$. To compute the variance over trajectories with $v_i(t) = v$ we bin trajectories by VAF size. Then a linear regression fit using weighted least squares (weighted by bin size) yields (Fig.3B):

$$\sigma^2_{v_i(t_0)=v}(\Delta_{v_i}(t)) \approx 2 \times 10^{-4} v. \tag{36}$$

4 Estimation of the number of HSPCs in participants of the LBC.

Combining (35) and (36) we can provide an approximation of the number of HSPCs in an individual

$$N_w \approx \frac{\lambda(1-2v_f(t))}{2 \times 10^{-4}}. \tag{37}$$

Denoting by v_i the mean VAF of a clone $v_i(t)$ over the observed time span in the LBC, we approximate

$$v_f(t) \approx \operatorname*{median}_{p \in \mathrm{LBC}} \max_{i \in I} v_i \approx 3.38 \times 10^{-2}, \tag{38}$$

as the median over all participants in the LBC of the mean VAF of their largest clone (SuppFig.2A). Further, we use the well-established estimate that HSPCs divide symmetrically for self-renewal every 40 weeks (Catlin et al., 2011), $\lambda = 1.3$, and estimate

$$N_w \approx 6000 \tag{39}$$

4.1 Estimations not accounting for the presence of fit clones

Since fit variants remain relatively small, notice that estimate (39) does not differ much from the estimate obtained directly from the evolution of neutral clones in the absence of a fit clone (20),

$$N_w = \frac{\lambda}{2 \times 10^{-4}} \approx 6500. \tag{40}$$

An even simpler approach to estimating the number of HSCs is provided by the distribution of neutral clones (16). At age t the maximum VAF reached by neutral clones, c_i, should be .

$$\begin{aligned} \max_{n_i(0)=1} v_i(t) &\approx \mu_{v_i}(t) + 2\sigma_{v_i}(t) \\ &= N_w^{-1}\left(0.5 + \sqrt{2\lambda t}\right). \end{aligned} \tag{41}$$

In the LBC data, we estimate $\max_{n_i(0)=1} v_i(t)$ using the mean size and age of synonymous trajectories v_s (SuppFig.2D):

$$\operatorname*{mean}_{p \in \text{LBC}} \max_{v_s \in p} v_s \approx 1 \times 10^{-2}. \tag{42}$$

Using the average age of participants during the observation of the trajectories used to compute the above estimate, $t \approx 80$ years, we obtain the rough estimate

$$N_w \approx 1000. \tag{43}$$

5 Drift-induced fluctuations filter

In the LBC, we estimate linear relations (30) and (36) and select clonal trajectories, $v(t)$, growing more than the expected growth over the observed time span,

$$\Delta_v(t) > \mu_{v_i(t_0)=v}(\Delta_{v_i}(t)) + 2\sigma^2_{v_i(t_0)=v}(\Delta_{v_i}(t)), \tag{44}$$

as mutations conferring a fitness advantage, $s > 0$.

Notice that, although this sieving method depends on the length of the observation period $t - t_0$, estimate (30) takes into account the average observation period of clonal trajectories. Further, the square root scaling of time in (27) reduces in the impact of the difference of observation periods; a span of 3 to 5 data points per trajectory yield $\sqrt{6} \leq \sqrt{t - t_0} \leq \sqrt{12}$.

6 Estimating clonal fitness

Given an individual with a collection of fit clones $\{c_i\}_{i\in F}$ with fitness parameters $\{s_i\}_{i\in F}$, their predicted deterministic time evolution is

$$v_f(t) = \frac{e^{s_f(t-t_f)}}{N_w + \sum_{i\in F} e^{s_i(t-t_i)}}, \tag{45}$$

for $f \in F$. We fit this evolution to data using least squares to infer parameters (s_f, t_f) for each clone $\{c_f\}_{f\in F}$ and a parameter N_w shared across all clones present in the individual.

We note the existence of a hypersurface of parameters yielding the same evolution. In the case of a single fit clone, $v_f(t)$, equation (45) can be re-parametrised as

$$v_f(t) = \frac{1}{2 + 2a(s)e^{-st}}, \tag{46}$$

with $a(s) = N_w e^{st_f}$. Therefore, any combination of parameters N_w and t_f satisfying $N_w = a(s)e^{-st_f}$ will produce the same VAF evolution $v_f(t)$ (SuppFig.2E and 2F). Let us highlight that, although different choices of N_w will result in a different t_f, the order of mutation acquisition remains unchanged with any choice of parameters in the specified hypersurface.

In this study we have refrained from fitting the 2-parameter model given both the impracticality of providing realistic bounds to $a(s)$ (constraining N_w within a pre-established ranges and $t_0 < t_f < t$ within achievable times as a function of fitness) and to show the existence of global solutions that span across different estimates of N_w in the literature through no or minor adjustments of inferred fitness values (SuppFig.2E and 2F).

7 Multiple fit clones

If two clones with $s_{1,2} > 0$ are present in an individual, the VAF is given by

$$v_i = \frac{e^{s_i t}}{N_0 + e^{s_1 t} + e^{s_2 t}}. \tag{47}$$

Assume without loss of generality that the second clone is fitter, $s_2 > s_1$, and consider under what conditions the VAF of the first clone will shrink, i.e.,

$$\frac{\mathrm{d}v_1}{\mathrm{d}t} < 0, \tag{48}$$

which is equivalent to

$$s_1 v_1 < \frac{e^{s_1 t}\left(s_1 e^{s_1 t} + s_2 e^{s_2 t}\right)}{\left(N_0 + e^{s_1 t} + e^{s_2 t}\right)^2}. \tag{49}$$

Solving for t gives

$$t > \frac{1}{s_2} \ln\left(\frac{s_1}{\Delta s} N_0\right), \tag{50}$$

where $\Delta s = s_2 - s_1$. From this we can see the fitter clone will always cause the first clone's VAF to shrink for times greater than (50). As we expect, this happens earlier for fitter secondary clones (greater s_2), and there is a further logarithmic dependence on the fitness of the first clone (later shrinking for greater s_1), the size of the neutral HSPC pool (later shrinking for greater N_0), and the difference in fitness between the two clones (earlier shrinking for larger Δs).

References

N. T. Bailey. *The elements of stochastic processes with applications to the natural sciences*, volume 25. John Wiley & Sons, 1990.

S. N. Catlin, L. Busque, R. E. Gale, P. Guttorp, and J. L. Abkowitz. The replication rate of human hematopoietic stem cells in vivo. *Blood*, 117(17): 4460–4466, 2011.

J. E. Till, E. A. McCulloch, and L. Siminovitch. A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proceedings of the National Academy of Sciences of the United States of America*, 51(1): 29, 1964.

C. J. Watson, A. Papula, G. Y. Poon, W. H. Wong, A. L. Young, T. E. Druley, D. S. Fisher, and J. R. Blundell. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*, 367(6485):1449–1454, 2020.