THE UNIVERSITY *of* EDINBURGH

THE UNIVERSITY *of* EDINBURGH

THE AGE-RELATED SOMATIC EVOLUTION OF CLONAL HAEMATOPOIESIS

Neil Alistair Robertson

Doctorate in the College of Medicine and Veterinary Medicine

Primary Supervisor: Tamir Chandra

Secondary Supervisors: Colin Semple and Thanasis Tsanas

Year of Presentation: 2023

Main Text: 24,345 Words

Abstract

Over a lifetime, human cells continually acquire mutations, some of which may alter cell division's complex homeostasis and lead to the subsequent expansion of somatic clones. Such expansions are frequent in the haematopoietic system and become detectable as we age.

Haematopoiesis is a complex and hierarchical system that generates millions of functionally diverse cells daily. This multi-tiered system allows for the rapid regeneration of our blood system in response to stress whilst protecting the pool of long-lived haematopoietic stem and progenitor cells (HSPCs) from excessive replicative stress. Haematopoiesis can function with high fidelity for many decades but is inevitably challenged by ageing and the time-dependent accumulation of somatic variation.

Clonal Haematopoiesis of Indeterminate Potential (CHIP) is defined as the expansion of HSPCs in healthy-aged individuals that results from genetic alterations. Although mostly inconsequential, the constant rate of the acquisition of mutations in HSPCs (17 mutations/year) leads to an increasing probability, with respect to age, of a variant occurring in a gene that dysregulates the tightly maintained mechanism of haematopoiesis. In healthy individuals, the differentiated cells that comprise our blood are the progeny of equally contributing stem cells that produce a genetically diverse, polyclonal population. CHIP, however, is marked by the population of blood cells becoming increasingly dominated by single (or multiple) genetic clone(s) that are genotypically identical.

In 2014, several independent studies [1, 2] confirmed that CHIP is a condition that increases with age: more than 10% of the population over 65 years are affected, with a prevalence that increases dramatically over subsequent decades. It has been associated with all-cause mortality, cardiovascular disease and haematological malignancies – a risk that scales with clone size [3].

Observing the relationship between CHIP and age-related pathologies, we sought to test the relationship between clonal haematopoiesis and ageing using a range of

published epigenetic clocks (which use DNA methylation states to predict biological age) to assess any association with biological ageing. In Robertson et al. [4], we characterised the landscape of somatic mutations in a range of core haematopoietic marker genes in the Lothian Birth Cohorts (LBCs). The LBCs are two parallel studies of ageing that consist of individuals over 70 and 79 years, in LBC36 and LBC21, respectively. We observed a significant association with biological ageing in several published cell-intrinsic clocks in participants that harboured a mutation in one of the six most prevalent CHIP genes versus our control group. CHIP status was associated with a significant increase in Horvath age acceleration: with an increase of 4.5 (SE 0.9) years in LBC1936 and 3.7 (SE 1.2) years in LBC1921 ($p = 2.3 \times 10\text{-}6$ and $2.5 \times 10\text{-}3$, respectively). In addition, we note significant epigenetic age accelerations in both DNMT3A and TET2 in isolation – the most commonly affected genes in clonal haematopoiesis. This result might indicate that CHIP is either driven or a driver of systemic ageing, explaining its links to non-haematological age-dependent pathologies.

A triptych of fundamental forces shape evolution: mutation, drift and selection. Whilst the first two are essentially stochastic processes, the third is the driving force: aiming to maximize fitness within an environment. Currently, it's not understood whether mutations in differing CHIP genes lead to distinct fitness advantages that would lead to patient stratification. Since mutations in HSPCs often instigate leukaemia, we hypothesize that HSC fitness substantially contributes to the transformation to disease states. We again leverage the LBCs, using a particularly unique aspect of their curation - the collection of peripheral blood over 12 years of later life - to develop a longitudinal assay for HSPC fitness using error-corrected sequencing. We then create a novel method we call the likelihood-based filter for time series data (LiFT) to determine fitness effects across our longitudinal data, quantifying the growth potential of somatic mutations within each participant. This approach discriminates naturally drifting populations of cells that typically harbour synonymous or non-functional variants and those that harbour driver mutations that give rise to rapidly growing clones. We characterise the fitness effects of mutations in many known CHIP driver genes and observe that differences in gene-specific fitness effects outweigh inter-

individual variation, which could constitute a new method of personalised clinical management [5].

This work has shown that CHIP confers a strong association with biological ageing through the prism of epigenetic clocks. Furthermore, we have begun characterising the fitness effects of genes that characterise CHIP whilst beginning to understand the molecular mechanisms behind their distinct fitness effects. We hope this should aid in a greater understanding of the pathogenesis of CHIP and assist in the improved stratification of patients.

Lay Summary

Clonal haematopoiesis is a condition that occurs when a single damaged blood stem cell starts to produce many identical copies of itself – known as a clone – that leads to an increase in the number of blood cells that carry the same mutation. This condition is relatively common in older individuals and can potentially increase the chances of developing blood cancers or other illnesses in later life. As we age, our blood system will acquire mutations, though only a handful will cause a cell to expand into a clone and produce a larger proportion of our blood.

We currently cannot predict which patients will develop blood cancers and when and how they develop them, meaning we cannot provide treatment because we lack a clear picture of which damaged clones grow fastest and pose the most significant risk. This thesis aims to understand which mutations grow faster than others, thereby increasing our potential to manage patients with large clone sizes in their blood.

Signed Declaration

I, Neil Alistair Robertson, declare that:

(a) the thesis has been composed by the student, and

(b) either that the work is the student's own, or, if the student has been a member of a research group, that the student has made a substantial contribution to the work, such contribution being clearly indicated, and

(c) that the work has not been submitted for any other degree or professional qualification except as specified, and (d) that any included publications are the student's own work, except where indicated throughout the thesis and summarised and clearly identified on the declarations page of the thesis.

Please see "Thesis Structure and Contributions" (pg. 8) and "Publications" (pg. 8) for full lists of publications that have been included in this thesis and contributions made to this body of work.

Neil A. Robertson

24th May 2023

Acknowledgements

I would like to thank my supervisor, Tamir Chandra, for being an incredibly supportive mentor over these last four years. He has created a truly liberating and creative space for his staff and it has genuinely been a fantastic place to work and develop. From the conferences we've attended together to the many nights at his house, your friendship has made the last few years relatively pain free. And to the members of the Chandra Lab, Nelly Olova and Daniel Simpson, thank you for making the early days of the lab so special.

I would also like to thank many of the additional supervisors who have helped me immeasurably over the last few years, including Kristina Kirschner, Eric Latorre-Crespo, Riccardo Marioni and Linus Schumacher. Your insights and wisdom have undoubtably helped me grow as a scientist and I thank you for all your patient advice. Undoubtably, so much of the work featured in this thesis would not have been possible without your diverse talents.

To my now wife, Amy, thank you for your patience these last few years. Hopefully this is the last of my most ridiculous and impoverishing schemes, although I look forward to many more! And of course, to my parents. As I said at the wedding, there is simply not enough space and time to thank you for everything. Thank you for the support, encouragement and love for all these years. I would not have been able to do this without you. I would also like to take this opportunity to my friends for their support over the years, in particular to David Calder and Scott Anderson for encouraging me to go through with the PhD at this point in my life.

Finally, I would like to thank my gran. Many years ago, I disappointed her by dropping out of medical school. At 102, she is still fiercely independent and her tenacity and strength are a constant inspiration. I would like to dedicate this to her.

Thesis Structure and Contributions

This work comprises an introduction chapter (Chapter 1), a methods section (Chapter 2), two results chapters (Chapter 3 and 4) and a discussion chapter (Chapter 5). Each results chapter is broken down into an introduction, results and conclusion section. My primary supervisor was Dr Tamir Chandra, with additional supervision from Dr Kristina Kirschner and Dr Linus Schumacher. My supervisory committee comprised of Dr Colin Semple and Dr Thanasis Tsanas.

The first chapter is a literature review that discusses the nature and theories of ageing and somatic evolution of cancer, before progressing to a discussion on the haematopoietic system, clonal haematopoiesis and epigenetic clock measurements. Chapter 2 describes the methods employed in Chapters 3 and 4.

Results chapter three was published in Current Biology (Robertson et al., 2019). Here I was first author. I conducted the majority of the analysis under the supervision of Dr Tamir Chandra, Dr Ricardo Marioni and Dr Kristina Kirschner. Many thanks to Ian Deary and the Lothian Birth Cohort team and participants for use of the data. See section a) below.

Results chapter 4 was published in Nature Medicine (Robertson and Latorre-Crespo et al., 2022). This was a truly interdisciplinary study comprising bioinformaticians (myself), ageing and haematology specialists (Dr Tamir Chandra and Dr Kristina Kirschner) and mathematicians and modellers (Dr Eric Latorre-Crespo (co-first author) and Dr Linus Schumacher). Here I performed the bioinformatic analysis and contributed to the figures, writing, conclusions and design of the study. See section b) below.

a) Robertson et al. (2019): T.C., R.E.M., K.K., and I.J.D. were involved in the conceptualization, funding acquisition and supervision of the study. R.E.M., T.C. and **N.A.R.** were involved in investigative and formal data analysis. T.C., R.E.M., K.K. and **N.A.R.** wrote the original draft. R.E.M., T.C., **N.A.R.**, R.F.H., D.L.M., M.T.T., J.H.,

D.S. and I.J.D. were involved in data curation. All authors reviewed and edited the manuscript.

b) Robertson and Latorre-Crespo et al., (2022): L.J.S., K.K. and T.C. conceived and supervised the study. **N.A.R.**, E.L.C., L.J.S., K.K. and T.C. wrote the manuscript. L.M., A.F. and L.M.G. generated data. **N.A.R.** and E.L.C. developed the methodology for data analysis. **N.A.R.**, E.L.C., M.T.T., A.C.P., J.A.M., B.J.L., J.L.P., R.F.H., R.E.M. and J.L.P. conducted data analysis. S.E.H., S.R.C. and I.J.D. curated the LBCs and gave access to samples. M.C. advised on aspects of the study.

Publications

Below is a list of publications that I have been personally involved with during my PhD. The first three are pertinent to this work and have been included in the appendix.

1. **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.," *Current Biology*, 29(16), pp. R786–R787. doi: 10.1016/j.cub.2019.07.011.

2. **Robertson, N. A.***, Latorre-Crespo, E.*, Terradas-Terradas, M., Lemos-Portela, J., Purcell, A. C., Livesey, B. J., Hillary, R. F., Murphy, L., Fawkes, A., MacGillivray, L., Copland, M., Marioni, R. E., Marsh, J. A., Harris, S. E., Cox, S. R., Deary, I. J., Schumacher, L. J., Kirschner, K. and Chandra, T. (2022) "Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects.," *Nature Medicine*, 28(7), pp. 1439–1446. doi: 10.1038/s41591-022-01883-3.

3. Terradas-Terradas, M., **Robertson, N. A.**, Chandra, T. and Kirschner, K. (2020) "Clonality in haematopoietic stem cell ageing.," *Mechanisms of Ageing and Development*, 189, p. 111279. doi: 10.1016/j.mad.2020.111279.

4. Gadd, D. A., Hillary, R. F., McCartney, D. L., Shi, L., Stolicyn, A., **Robertson, N. A.**, Walker, R. M., McGeachan, R. I., Campbell, A., Xueyi, S., Barbu, M. C., Green, C., Morris, S. W., Harris, M. A., Backhouse, E. V., Wardlaw, J. M., Steele, J. D., Oyarzún, D. A., Muniz-Terrera, G., Ritchie, C., Nevado-Holgado, A., Chandra, T., Hayward, C., Evans, K. L., Porteous, D. J., Cox, S. R., Whalley, H. C., McIntosh, A. M. and Marioni, R. E. (2022) "Integrated methylome and phenome study of the circulating proteome reveals markers pertinent to brain health.," *Nature Communications*, 13(1), p. 4670. doi: 10.1038/s41467-022-32319-8.

5. Guzniczak, E., Otto, O., Whyte, G., Chandra, T., **Robertson, N. A.**, Willoughby, N., Jimenez, M. and Bridle, H. (2020) "Purifying stem cell-derived red blood cells: a high-throughput label-free downstream processing strategy based on microfluidic spiral inertial separation and membrane filtration.,"

Biotechnology and Bioengineering, 117(7), pp. 2032–2045. doi: 10.1002/bit.27319.

List of Abbreviations

5hmC	5-hydroxy-methyl-cytosine
5mC	5-methyl-cytosine
AML	Acute myeloid leukaemia
ARCH	Age-related clonal haematopoiesis
ASXL1	Additional Sex Combs Like Transcriptional Regulator 1
bp	Base pair
CBL	CBL proto-oncogene
chAge	chronological age
CpG	5'—C—phosphate—G—3'
CTCF	CCCTC-binding factor
CH	Clonal haematopoiesis
CHIP	Clonal haematopoiesis of Indeterminate Potential
CLL	Chronic lymphocytic leukaemia
CML	Chronic Myeloid Leukaemia
COPD	Chronic obstructive pulmonary disease
DNA	Deoxyribonucleic acid
DNAm	DNA methylation
DNMT3A	DNA Methyltransferase 3 Alpha
DDR	DNA damage response
eAge	Epigenetic age
EEAA	Extrinsic epigenetic age acceleration
FLT3	FMS Related Receptor Tyrosine Kinase 3
GOF	Gain of Function
HDAC	Histone deacetylase
HSC	Haematopoietic Stem Cell
HSPC	Haematopoietic Stem and Progenitor Cell
IEAA	Intrinsic epigenetic age acceleration
IFNα	Interferon alpha
IFNγ	Interferon gamma
IGF	Insulin-like growth factor

JAK	Janus kinase 2
LBC	Lothian Birth Cohort
LBC1936	Lothian Birth Cohort of 1936
LBC1921	Lothian Birth Cohort of 1921
log2	logarithm base 2
log10	logarithm base 10
LASSO	Least absolute shrinkage operator
LiFT	Likelihood-based filter for time series data
LOF	Loss of Function
LOY	Loss of Y-chromosome
mCA	Mosaic chromosomal alteration
MPN	Myeloproliferative Neoplasms
NGS	Next Generation Sequencing
NPM1	Nucleophosmin 1
PBMC	Peripheral Blood Mononuclear Cell
PMF	Primary Myelofibrosis
PPM1D	Protein phosphatase Mg2+/Mn2+-dependent 1D
PV	Polycythaemia Vera
RNA	Ribonucleic Acid
SF3B1	Splicing Factor 3b Subunit 1
SRSF2	Serine and Arginine Rich Splicing Factor 2
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TET2	Tet Methylcytosine Dioxygenase 2
TP53	Tumour Protein P53
U2AF1	U2 small nuclear RNA auxiliary factor 1
VAF	Variant Allele Frequency
WGS	Whole Genome Sequencing
WT	Wild-Type
XCI	X-Chromosome Inactivation

List of Figures

List of Tables

Table of Contents

Chapter 1: Literature Review

1.1 The Haematopoietic System

Haematopoiesis is the process that creates the cellular components of our blood. For perspective, over one trillion blood cells develop every day from the haematopoietic stem cells in our bone marrow, generating a functionally diverse range of mature cells [6]. In achieving this complexity, haematopoiesis requires a "hierarchy of progenitors" whose fates become increasingly limited within their lineages [7]. Haematopoietic stem cells (HSCs) reside at the root of this heterogenous system and serve two key roles: 1) to replenishing the HSC pool through repeated cycles of self-renewal (where a dividing HSC produces two identical HSCs), and; 2) generating pluripotent myeloid or lymphoid cells through asymmetrical division [8]. This complex hierarchy provides a pool of progenitor cells to be called upon under a varied range of stressors – such as infection or blood loss - while protecting against excessive replication at the top of the lineage [7].

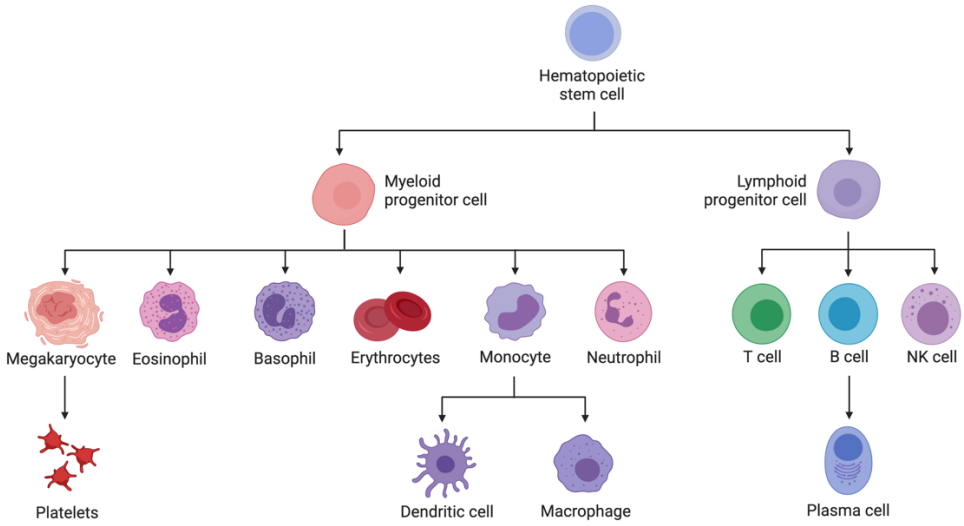


Figure 1.1: Haematopoietic hierarchy in normal conditions. *Depicting the conventional hierarchical structure of haematopoiesis. Haematopoietic stem cells (HSCs) commit to their lineage through a process of differentiation driven by a myriad of cell-type specific internal and external stimuli.*

It has been noted that our blood system emerges from a single cell at gastrulation [9]. Thereafter, early haematopoietic function begins in the yolk sack, then the spleen and liver at around 80 days after conception before appearing in the bone marrow as we get closer to birth [10]. We are born with an estimated 50,000 to 200,000 HSCs that are responsible for constituting the blood system across our lifespan – a remarkably small number given the level of productivity and output required over this period [9]. From the perspective of somatic evolution, this demarks the origin, period and boundaries of which a single germline genotype can create the haematopoietic system and of which somatic variation can feasibly affect this system.

Haematopoiesis allows HSCs to differentiate into a landscape of progenitor cell types and can be best thought of as the cascading reduction of multilineage potential into greater functional specification [11, 12]. The gradual identification of cell surface markers has illuminated our understanding of the relative populations and lineage potentials of haematopoietic cells. Protein markers that typically denote this specification have been labelled as “clusters of differentiation” (CD) and include CD34, CD38, CD4 and CD8 amongst others. HSCs typically express CD34 without lineage (Lin) markers and are classed as $Lin^-/CD34^+/CD38^-$ and as they commit to differentiation, they begin to express CD38 in early progenitor cells (as $Lin^-/CD34^+/CD38^+$). Thereafter, they can present a complex landscape of lineage specific (Lin^+) cell-surface markers as they gradually commit to lineage specificity [12].

HSCs form the foundation of the haematopoietic system and can differentiate into two distinct myeloid and lymphoid lineages (Figure 1.1). Simplistically, the lymphoid lineage comprises several progenitor subtypes that migrate to the thalamus, spleen or lymph nodes to complete their differentiation into T-, B- or natural killer (NK) cells, respectively. Myeloid progenitors are a morphologically diverse grouping that consists of basophils, eosinophils and neutrophils alongside a monocytic lineage that contributes macrophages and dendritic cells (DCs) and megakaryocytes – the precursor to platelets [13, 14]. Together, this complex complement of cell types allows our blood system to perform its key roles in innate and adaptive immunity, blood clotting and nutrient transport.

The continuous population of our blood system requires rigorous control of decisions regarding cell fate. The choice and timing of decisions that govern self-renewal or differentiation, or quiescence and proliferation, amongst others, requires input from a multitude of sources including varied transcription factor activity, epigenetic changes in regulatory regions (in particular transcription factor binding sites) and external stimulus from the niche [6, 7, 15]. These changes evoke patterns of gene expression that gradually changes cell identity as it progresses through a lineage and have been studied extensively using single-cell RNA-sequencing [16–20] and even probabilistic methods [21, 22]. Transcription factors guide their own lineage specification while acting to overrule others that favour different trajectories – the combination of positive and inimical roles in transcription factor function help to reinforce lineage commitment [23–25].

The bone marrow (BM) microenvironment provides the environs for the sustained maintenance and function of HSCs. It plays an important role in the endogenous signalling controlling the mobilization, regulation and cell-fate decisions of HSCs [15, 26]. HSCs are thought to exist of two main compartment types – the endosteal osteoblastic niche and the perivascular endothelial niche – that are functionally heterogeneous and are thought to contribute to distinct subsets of HSCs [15]. The BM niche comprises a varied array of cell-types; including but not exclusively, fat, endothelial, vascular and osteoblast derived cells; that are each thought to contribute differing extracellular cues. These external inputs can prompt a range of outcomes that promote heterogeneity in the self-renewing capacity of HSCs, or can trigger differentiation as a response to inflammation, infection or other systemic stimuli [27]. While time provides many intrinsic insults to HSCs through sustained DNA damage, changes in the bone marrow niche are an important factor affecting the haematopoietic system as we age.

1.1.1 Ageing in the Haematopoietic System

Ageing is a slow process and many of its associated phenotypes can appear ambiguous across this long timeframe. However, ageing inexorably leads to the

decline of physiological systems and is the most important risk factor in the development of many important pathologies, such as cancer, neurodegeneration and cardiovascular disease [28].

The role ageing plays in the haematopoietic system is multifaceted. Firstly, and somewhat counterintuitively, the number of HSCs increases with age in both humans and mice. However, this increase in population size does not translate with sustained or increased functionality [29, 30]. This enlarged HSC pool has increased self-renewal and reduced regenerative capacity leading to a net loss of function [31].

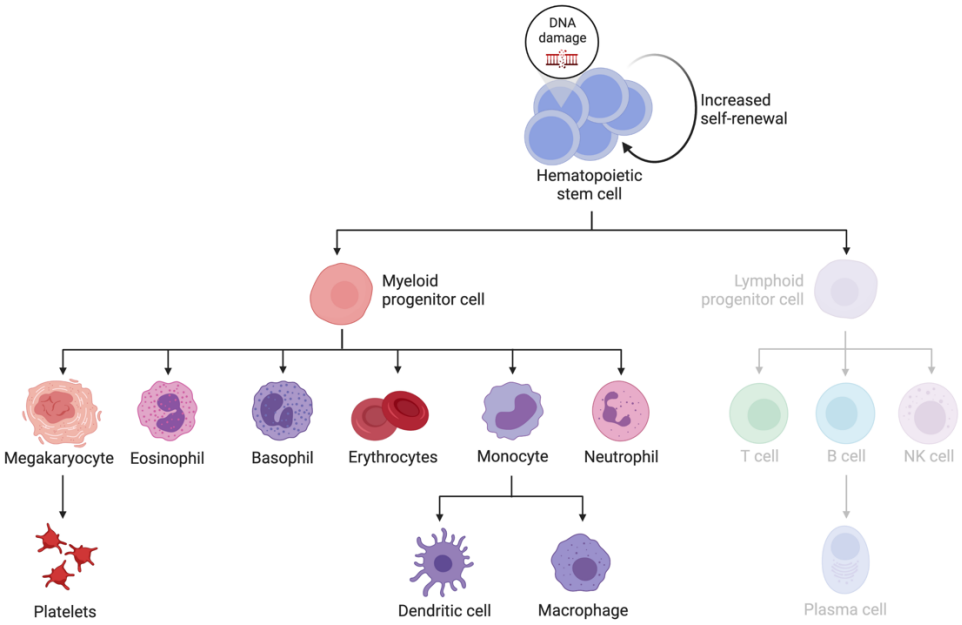


Figure 1.2: The haematopoietic system with age. *Accumulation of DNA damage, changes in the systemic and niche signalling profiles and the erosion of epigenetic marks leads to increased self-renewal, homogenisation of the stem-cell pool and myeloid lineage biases.*

The aged haematopoietic system exhibits skewed differentiation. Typically, this results in a marked reduction in the output of lymphoid and erythroid cells, with cells in the myeloid lineage displaying stable or even increased outputs [32, 33]. The reduction in lymphoid cells can result in immunodeficiencies and are thought to be driven by the homogenisation of the stem cell pool and through sustained extracellular signalling

that exhausts the immune/inflammation directed lymphoid biased HSC population [34, 35].

Changes in the relative concentrations of systemic factors and bone marrow niche signalling also contribute to HSC ageing [36]. The gradual loss of Insulin Growth Factor 1 (IGF1) signalling that begins in middle age has been shown to lead to many of the hallmarks of HSC ageing in mice, including mitochondrial disfunction and impaired differentiation potential [37]. While conventional inflammatory signalling plays an essential role in triggering proliferation and the immune-response, sustained age-dependent inflammatory signalling eventually results in exhaustion of the stem cell pool [38].

HSCs, like all cells, can become victims of typical age-related impairments like increased DNA damage, shortened telomeres and loss of epigenetic fidelity. The outcome of these lesions, coupled with the changes to extrinsic signalling profiles, lead to impairments in the haematopoietic system that typically resemble the beginnings of haematological malignancy (Figure 1.2) [36].

1.2 Clonal Haematopoiesis

Clonal haematopoiesis (CH) - or as it has hitherto been described within the clinic as clonal haematopoiesis of indeterminate potential (CHIP) – is defined as the clonal expansion of haematopoietic stem and progenitor cells (HSPCs) in healthy aged individuals. Although mostly inconsequential, the constant rate of acquisition of mutations in HSPCs (17 mutations/year [39]) leads to an increasing probability, with respect to age, of a somatic variant occurring that can destabilise the tightly regulated homeostasis of haematopoiesis (Figure 1.3). In healthy individuals, differentiated blood cells are the net progeny of an approximately balanced HSPC pool and together produce a well-mixed pool of differentiated cells without any single mutations reaching high variant allele frequencies (VAFs). Clonal haematopoiesis, however, is marked by the population of blood cells showing increasing oligoclonality through selection - becoming increasing dominated by single (or multiple) large genetic clones that are genotypically identical.

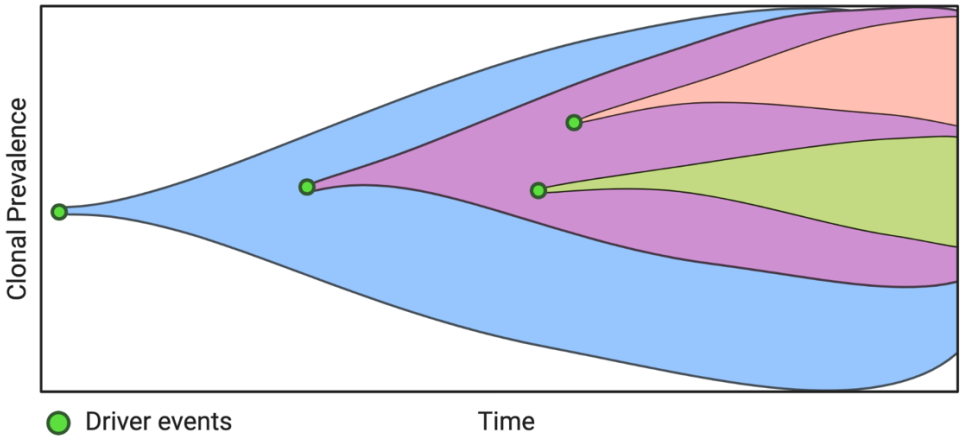


Figure 1.3: Schematic describing the effects of CH driver mutations on the stem cell pool. *With time, the effect of positive selection caused by mutations in specific gene drivers increases the oligoclonal burden of the haematopoietic system, with multiple occurring hits increasing the potential for malignant transformation.*

Here we will discuss the background and recent advances within the field and give and overview of the current perspective on the cause and consequences of clonal haematopoiesis.

1.2.1 Early Insights into Haematopoietic Oligoclonality

Thirty years ago, the first observations of clonality in the haematopoietic system were made when a number of studies showed maternal to paternal biases in X-chromosome inactivation (XCI) [40–42]. While XCI is a normal developmental process that maintains an equilibrium of X-linked gene activity across XX females and XY males, early work by Gale and others uncovered an increase in XCI activity in healthy woman that appeared to correlate with age and changes in blood cell-type compositions, notably myeloid biases [42]. While the causal mechanisms remained unclear, the eventual confirmation of acquired non-random X-chromosome inactivation in females was considered to be early evidence of stochastic clonal skewing of haematopoietic output with respect to age [43]. Subsequent sequencing of individuals with XCI skewing (and their associated lineage biases) showed an enrichment for mutations in the *TET2* gene, a known modulator of methylation states, suggesting a genetic basis for this phenomenon [44].

Following these early observations, a greater understanding of the somatic mutational landscape of common blood cancers allowed researchers a more comprehensive understanding of the genetic variants that drive these pathologies. The traditional multistep model of tumorigenesis predicts the step-wise loss of key tumour suppressor genes alongside the over-expression of oncogenes that will result in the eventual transformation to cancer. Such models, developed by Bert Vogelstein and others, were formulated by observing the over-representation of genetic alterations – in colorectal cancer, the *APC* gene is most commonly altered and is thus considered a foundational step in cancer formation [45]. It is through this prism that the genetic basis of clonal haematopoiesis was discovered.

Analogous to many other cancers, AML develops in a time-dependent manner through the sequential accumulation of driver mutations [46]. If an early genetic lesion can drive a clonal expansion without developing to cancer, presumably it would be possible to isolate a premalignant stage where only the founding mutation is present [47]. These earliest studies involved genetic sequencing of AML sufferers and unravelled

several key aspects that confirmed clonal haematopoiesis as a generalisable origin of the pathology: firstly, several studies observed similar enrichments of the presumed founding mutations, including *TET2* [48] and *DNMT3A* [49] with associated enrichments for particular pathways including methylome and chromatin modifiers. Secondly, on the timing and order of mutation acquisition: groups began to compare the mutational burden between AML and skin controls which showed similar counts of passenger variants between the two tissues suggesting the stochastic time-dependence of their acquisition. Then, that in HSPCs, the "*number of additional passengers added with progression events is typically much smaller than the number of passengers captured with the initiating event (which accumulated over the lifetime of the founding cell)*" [46] suggesting that lesions that induce a clonal expansion also carry the genetic history of the clone and can be ordered accordingly. Alongside this, others noted the size of the clonal population carrying a given progression event may correspond to increased self-renewal against a background of mutations that confer no advantage and are gradually lost to drift [50].

These experiments highlighted the ability of genomic sequencing to capture both the targets, history and temporal ordering of mutations and identified an initiating state of increased HSPC self-renewal that occurs years before pathological relevance. Despite this, a full understanding of the mutational spectrum, aetiology and the outcomes associated with clonal haematopoiesis in the general population remained unknown.

1.2.2 Clonal Haematopoiesis in the Next Generation Sequencing Age

Beginning around 2014, several teams began to independently investigate exome sequencing data from several cohorts comprising over 30,000 individuals for somatic variants that might be associated with haematological disease [1, 2, 51, 52]. These studies can be considered to be at a "population level" as they comprised participants from a range of age-groupings – generally from early adulthood to late life – and covered what can be considered the normal burden of disease in a population without selecting for specific haematological phenotypes. Most importantly, the source of the

studied genomic data was peripheral blood – allowing them to study the mutational spectrum of clonal haematopoiesis to an extent that had previously been impossible. Some have described these studies as an experiment in "saturation mutagenesis" [53]: in a population of sufficient size, all possible mutations that can occur *will* occur in HSPCs; ergo, mutations that are damaging or neutral (that carry no discernible or negative fitness advantages) will not result in clonal growth that will lead to their detection in the blood and will be lost to genetic drift. Any mutation that can be detected can be presumed to be a driver of clonal haematopoiesis and will point to biological pathways that increase the fitness of HSPC growth.

Despite participants in each study coming from several distinct populations and cohorts it was initially counterintuitive to find that clonal haematopoiesis is driven by variants in a small subset of genes [54–56]. While mutations in canonical tumour suppressors and oncogenes regularly observed in cancer were present - such as those associated with DNA damage response, growth and survival signalling – nearly two thirds came from just two genes involved in the maintenance of DNA methylation, *DNMT3A* and *TET2*. Behind this, *ASXL1*, a chromatin modifier was the third most heavily affected, alongside a range of splicing factors (SRSF2, U2AF, SF3B1). Why simple loss of function (LOF) mutations in these genes induces the clonal expansion of HSPCs remains an open area of research.

Another notable finding from these early cohort studies suggested that the incidence of clonal haematopoiesis is age-dependent and increases in frequency across a lifespan. In individuals under 40 years of age, the burden of clonal haematopoiesis is less than 1%. However, this burden escalates to around 10-20% of individuals above the age of 70 years and can even achieve a prevalence of nearly 70% by the age of 90 years depending on the sensitivity of the sequencing method employed [53, 54, 57]. Indeed, if we assume that circulating PBMCs are a representation of the somatic mutational burden of the HSPC population, our potential to detect CH within an individual is likely most readily affected by the accuracy and sampling potential of the sequencing platform. In these initial cross-sectional studies, the size of the detectable clones was large; a median of 18% of cumulative PBMCs carried mutations [1]. Through the use of exome-sequencing, we can derive a solid understanding of the

age-dependence and scale of the oligoclonal burden, however, a lack of sensitivity limits our ability to detect small clones (particularly at earlier age points) and highlights the importance of improved sequencing methods to accurately estimate the growth and origin of low frequency variants.

Mutational signature analysis – which attempts to classify the forms of intrinsic and extrinsic mutagenic processes that give rise to a specific pattern of genome-wide nucleotide switching due to somatic variation [58, 59] - identifies two specific patterns in clonal haematopoiesis: predominantly signature 6 (C>T), an age-associated signature; and signature 4 (C>A), commonly associated with exposure to smoking [1]. The enrichment of signature 2 (~60% or all classified mutations in the Jaiswal cohort) is purported to be driven by the endogenous and spontaneous deamination of methylcytosine and indicates that the mutational processes that drive clonal haematopoiesis likely arise through an age-dependent, although not necessarily linear, acquisition of somatic variation over a lifespan [60].

Many of the cases of CH describe mutations in genes that are known drivers of haematological disease (DNMT3A, TET2, ASXL1, JAK2, P53), yet despite the high burden of somatic clones in the elderly, the risk of progression to cancer was limited (1% per year) despite over 42% of blood cancers within the cohort displaying some form of CH driven clonality [2]. Clonal haematopoiesis was also shown to be associated with both cardiovascular disease and other distal pathologies of ageing [1, 39, 61]. However, these associations are still a matter of some debate [62, 63] and are potentially confounded by covariates linked to the over-arching processes of ageing where the incidence of CH is most abundant. Greater clarity on the complex associations with CH and secondary pathologies will likely become increasingly clarified by the next generation of population level cohort studies, like the UK Biobank, and will be discussed in more detail in later sections.

1.2.3 The Genes and Genetics of Clonal Haematopoiesis

Clonal haematopoiesis is the consequences of the outgrowth of high fitness clones that are generally driven by somatic mutations in a small set of functionally diverse

genes. Although captured under this nomenclature due to their capacity to permit progenitor and stem cell expansion in the haematopoietic niche, their diverse roles lead to a variety of differing mechanisms that permit this expansion with corresponding divergencies in their aetiology, pathophysiological outcomes and associated disease risk [63]. This thesis examines clonal haematopoiesis through the prism of a set of well described drivers of clonal haematopoiesis (Figure 1.4). Here, I attempt to provide short summaries and a discussion of their roles in positive selection and their downstream consequences. I refer to gene names in the *italic* format and proteins as unitalicized text (i.e., *DNMT3A* encodes for the protein, DNMT3A).

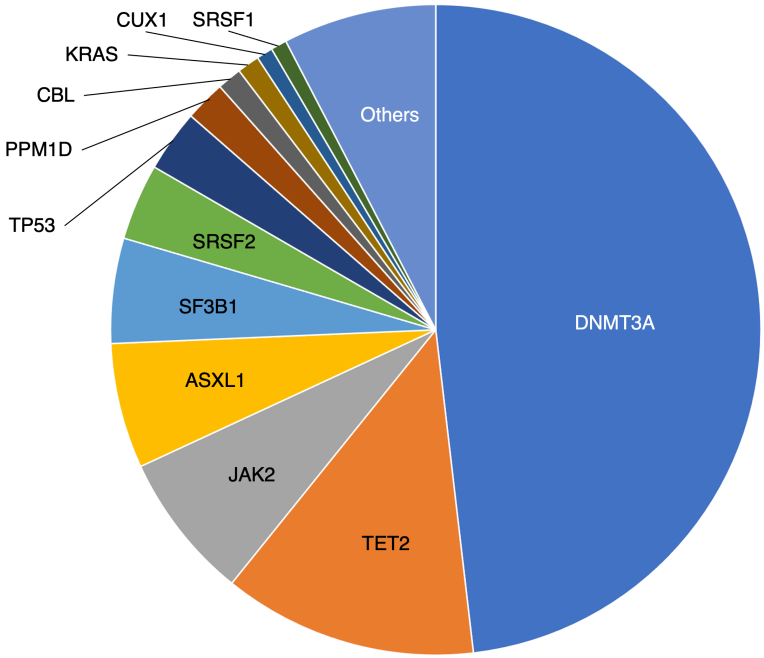


Figure 1.4: The pan-cohort prevalence of gene driver mutations. *The average proportions of the top 12 most prevalent genes harbouring known driver mutations in CH* [1, 2, 52, 54, 64]*. The "Others" category comprises 19 additional rarely observed genes. Cohorts were chosen due to lack of prior therapeutic or pathological selection criteria. As it is difficult to accurately correct for the variety of sequencing methods, age distributions or cohort selection methods, this plot is mainly illustrative of the canonical enrichment patterns of the genetic drivers in CH.*

The types of genes regularly mutated in CH tend to fall into several ontological classes: epigenetic and methylation regulators, splicing factors and genes involved in the DNA damage response; with less frequent numbers in developmental transcription factors, cohesion complex components and mitogenic regulators. Clonal haematopoiesis has

a complex and diverse genetic background and even minor changes to the genetic architecture of HSPCs can create conditions that can lead to significant effects when exhibited over many decades. In the section below, I attempt to devote time to the key genes that play the greatest contribution to CH and are of particular interest to this thesis, with an attempt to focus on the mechanisms that give rise to their expansion.

1.2.3.1 Regulation of DNA Methylation

DNA (cytosine-5)-methyltransferase 3A (*DNMT3A*)

DNMT3A encodes for one of the two key *de novo* methyltransferases, DNMT3A (alongside DNMT3B) whose role is to establish DNA methylation patterns across the genome [65]. DNA methylation is a major epigenetic regulator that has a vital role in X-chromosome inactivation, genomic stability [66], regulating gene expression [67] and cell differentiation [68]. DNA methylation is the addition of a methyl group (CH_3) to position C-5 in cytosine within the symmetric 5'—C—phosphate—G—3' (CpG) dinucleotide creating a methylated cytosine (5mC). More discussion will be made on the role of DNA methylation in later sections when I discuss its use in epigenetic estimates of age.

DNMT3A is highly expressed in haematopoietic stem and progenitor cells (HSPCs) [69] as well as during neurogenesis. Studies of DNMT family knock-outs in mouse models have shown these genes to be essential to development [68]. The role of DNMT3A in HSPCs is primarily to protect stem cells from excess multipotency - preserving the stem cell pool [70] – alongside orchestrating the complex and multi-layered process of haematopoiesis by manipulating gene expression programs via a repatterning of DNA methylation at key regulatory loci [71].

In HSPCs, mutations in *DNMT3A* have been shown to inhibit differentiation programs, whilst promoting self-renewal [72, 73], thereby reducing (or obstructing) multilineage potential. This occurs via loss, or erosion, of DNA methylation profiles at genes that promote stem-ness over differentiation – becoming increasingly reinforced through each cycle of self-renewal [74]. Coupled with this, DNMT3A loss has been shown to immortalise HSPCs [75] leading to skewed division potential. Taken together,

mutations in *DNMT3A* can preserve stem-ness and replicative potential [75, 76], permitting *DNMT3A* mutant cells to expand in the haematopoietic niche through repeated cycles of self-renewal, allowing it to outcompete WT HSPC equivalents over long periods in a cell-intrinsic manner.

Despite their apparently benign phenotype, high representation and assumed tolerance of *DNMT3A* mutant clones across numerous cohort studies [1, 2, 51, 52, 54, 55], the evidence suggests that DNMT3A has an important role in pre-leukemic disease [77]. This initially appears counterintuitive, however, their progression towards leukemic states likely requires cooperation from additional oncogenic mutations that hijack their resilience and pervasiveness [78, 79]. Mutations in *DNMT3A* are strongly linked with Acute Myeloid Leukaemia and between 20-30% of all sufferers are carriers for LOF mutations in this gene [49, 80], while *DNMT3A* mutations can also be seen as a strong predictor of AML onset – particularly in the young – with a risk that scales with clone size [3]. DNMT3A is regularly altered with other oncogenes: almost 30% of AML sufferers with mutations in DNMT3A have additional mutations in FMS Related Receptor Tyrosine Kinase 3 (*FLT3*) and Nucleophosmin 1 (*NPM1*) [81, 82] alongside a host of mutations that are more common in solid tumours, such as NRAS (G12V) [74]. Indeed, the persistence of *DNMT3A* mutated clones, might be a key factor in the difficulties in treating many leukemic diseases, due to their capacity to maintain stemness and resistance to chemotherapies while enduring long into remission [78, 79].

In the last few years, some studies have shown that chronic infection has been shown to increase the fitness of *DNMT3A* clones [83]. Chronic exposure to pathogens result in the long-term elevation of inflammation and immune responses and gradually exhaust the stem cell pool [84]. The subsequent selective pressure favours the expansion of pro-stemness mutant *DNMT3A* HSC clones against wild-type competitors [85] and highlights how cell intrinsic and extrinsic factors can interact to promote somatic mosaicism.

Tet Methylcytosine dioxygenase (*TET2*)

The second most frequently mutated gene in CH, *TET2*, is one of the ten-eleven translocation (TET) family of enzymes primarily responsible for the regulation of gene expression through modulation of methylation patterns [86, 87]. Not so long ago, DNA methylation was thought to be an irreversible state, until the discovery of its sister protein TET1 in 2009 [88]. TET2 and its sister proteins act to remove methyl groups from CpG dinucleotides via the oxidation of 5-methylcytosine (5mC) to form 5-hydroxymethylcytosine (5hmC), leading to a net hypomethylation effect across the methylome.

The TET2 protein is a large (~200kDa) multidomain enzyme. Mutations in TET2 tend to be heterozygous and loss-of-function resulting in aberrant methylation patterns that alter gene expression programmes. This enzyme is frequently mutated across a range of haematopoietic and solid cancers [89, 90]. *TET2* is widely transcribed in haematopoietic cell populations, from progenitors to mature lymphoid lineages [91]. Deletion of *TET2* in murine models has been shown to increase Lin–Sca-1+c-Kit+ (LSK) cell population – boosting haematopoietic repopulation via increased self-renewal potential, while biasing towards monocyte, macrophage and other myeloid lineages before progression to malignancy [92, 93]. TET2 also has a necessary role in the function of innate immunity and plays an important role in the regulation of IL6 – Interleukin 6; a mediator of inflammation – through the recruitment of histone deacetylases (namely HDAC6) to this locus [94].

TET2 is the second most prevalent mutation in clonal haematopoiesis (Figure 1.4) with a mutational prevalence that is largely age-dependent [1]. Several studies have now explored the cell-intrinsic mechanisms that lead to malignant transformation that result from *Tet2* knockouts in mice [93, 95, 96]. In each of these studies, an expansion of the HSC compartment was shown alongside increased self-renewal potential with loss of *TET2*. Some studies have shown that *TET2* deficiency can result in increased mutagenicity: Pan et al., have shown that this can lead to the spontaneous development of a range of differing haematological malignancies [90] and targeted single-cell sequencing from the same group revealed an increased mutation rate in *Tet2*-/- HSPCs, particularly in sites with elevated levels of 5-hydroxymethylcytosine, indicating malignant transformation can be driven via cell-intrinsic mechanisms [92].

Clonal haematopoiesis has well described links to inflammation and *TET2* likely plays a key role in this axis. While associations with CH and atherosclerosis have been shown [39, 97], Jaiswal and colleagues have also shown an acceleration in the development of atherosclerosis in mice with Tet2 knockout bone marrow transplantations in the presence of a high-cholesterol, high-fat diet [1]. *TET2* knockout macrophages, cultured with low-density lipoproteins have also shown marked increases in inflammatory transcriptional signatures compared to their wild-type (WT) comparators, suggesting a requirement for inflammatory signalling to progress atherosclerosis development in the *TET2* mutant background [98]. Despite this, the complex role inflammation plays in the growth of mutant *TET2* clones and subsequent pathologies remains unclear.

To date, *TET2* is the only gene shown to have any hereditary predisposition through the presence of a noncoding variant at an enhancer - a distal gene regulatory element - associated with expression of this enzyme [99].

TET2 and DNMT3A: Antagonistic Functions, Overlapping Phenotypes

It is still a matter of some debate as to why these two genes with such dichotomy of function results in similar phenotypes that give rise to increased self-renewal patterns and subsequent clonal expansion. Part of the reasoning behind the similarities between mutant *DNMT3A* and *TET2* clones are likely: a) they are biologically rather benign with limited pathological effects – their near ubiquitous presence in populations likely indicates low risk to cancer progression in isolation [100]; b) their rate of growth is slow with limited proliferative advantage, therefore, progression to a dominant position in the haematopoietic system likely takes decades and is rarely met in individuals [5, 101], and; c) despite their opposing nature in maintaining DNA methylation, the dynamic temporal mechanisms that underpin the role of DNA methylation are still poorly understood – clearly both genes are required to orchestrate the gene expression programs that are required to transition through the haematopoietic lineages, despite the poor correlation of DNA methylation patterns and gene expression between these enzymes [102, 103].

1.2.3.2 Histone Regulation

Additional Sex Combs-Like 1 (ASXL1)

The *ASXL1* gene encodes a protein that plays a role in epigenetic regulation and transcriptional repression through its interactions with chromatin modifiers and transcription factors through its involvement in the polycomb repressive complex (PCR) [104]. Mutations in this gene are frequently detected in both clonal haematopoiesis and many haematological cancers, such as myelodysplastic syndrome (MDS), chronic myeloproliferative neoplasms (MPN) and acute myeloid leukaemia (AML) [105]. Functionally, ASXL1 is responsible for mediating PRC2 histone methylation (primarily of histone H3) and polycomb repressive complex 1 (PRC1) as well as the deubiquitination of histone H2A [106, 107]. Mutations in *ASXL1* markedly deplete repressive mark H3K27me3 at PRC2 developmental genes that regulate stem cell homeostasis [108].

An analysis of the UK Bio Bank has suggested that carriers of *ASXL1* mutations tend to have a significant association with smoking history, suggesting that the growth of *ASXL1* mutant clones might have some dependence on increased inflammation or mutagenicity [109]. Indeed, mutant *ASXL1* HSPC clones have been shown to express anti-inflammatory factors that confer resistance to the inflammatory environs produced by their own progeny – exemplifying the levels of somatic evolution that are sometimes required to promote clonal growth [110].

Alongside *ASXL1*, several other epigenetic regulators have been detected in enriched quantities in populations of CH carriers. Histone methyltransferases such as EZH2, KMT2A and KDM6A have been discovered in small numbers, alongside other proteins involved in regulating the polycomb repressive complex, such as BCOR and BCORL1, suggesting overlapping functionality in promoting the expansion of HSPC clones.

1.2.3.3 Mitogenic Regulators

Janus Kinase 2 (JAK2)

JAK2 is a non-receptor tyrosine kinase that plays a vital role in cytokine signalling and regulation of growth factors that include erythropoietin, thrombopoietin and interleukin-3 (IL3) [111, 112]. The Janus Kinase family of proteins serve as the cytoplasmic signalling component mediating membrane to nuclear communication in cells - most notably through the JAK-STAT pathway [113, 114]. JAK-STAT signalling has a diverse array of functions, including; the orchestration of adaptive and innate immunity, cell-death, inflammation and haematopoiesis [114].

JAKs are unusual in comparison to other tyrosine kinases through their possession of a psuedokinase domain upstream of their functional tyrosine kinase locus which serves to maintain a stable and low basal level of kinase activity [115]. The mutational spectrum of *JAK2* in haematopoietic disorders is simple and relatively unique, with specific gain-of-function "hotspot" mutations targeting the psuedokinase domain that constitutively activates the signalling pathway, such as the JAK2 V617F mutation [116]. The JAK2 V617F mutation has been indicated as a key driver of clonal haematopoiesis and an array of haematological disease through its ability to generate a pro-proliferative phenotype through the downstream activation of JAK-STAT, RAS/RAF or P3K/AKT pathways [117].

Mutations in *JAK2* have been found in significant proportions of patients with MPNs, including 95% of cases of polycythaemia vera (PV) linked to the overproduction of red blood cells (RBCs) and around 50% of the cases of essential thrombocythemia (ET) which is associated with platelet overproduction [118–120]. Haematological malignancies likely arise when a higher clonal burden of mutant JAK2 is attained which can account for changes is constituent blood count proportions – or cytopenia's – in circulating blood [121]. Patients with JAK2 V617F that have a disease burden - including MPN subtypes, PV and ET - can be specifically differentiated from CHIP through measuring blood cell type proportions, such as haemoglobin, leukocyte and platelet counts, with sufferers frequently displaying higher mutational variant allele frequencies (VAFs) [122]. JAK2 V617F has also been shown to exacerbate cardiovascular disease through increased cytokine activity in mice [123].

Aside from *JAK2*, a long tail of infrequently mutated mitogenic proteins includes *JAK3*, *NRAS* and *KRAS* genes, alongside *NF1* and *PTPN11* - frequently enriched oncogenes in many solid cancers.

1.2.3.4 Spliceosomal Mutations

Mutations in splicing factors emerge late in the pathogenesis of clonal haematopoiesis with genes such as Splicing Factor 3b Subunit (*SF3B1*), Serine and Arginine Rich Spicing Factor 2 (*SRSF2*) and U2 Small Nuclear RNA Auxiliary Factor 1 (*USAF1*) being commonly affected [124]. RNA splicing involves the post-translational regulation of gene expression via the cleavage of intronic DNA at conserved sequence motifs known as splice sites [125, 126] and in turn radically increases the functional proteomic repertoire by allowing variation in exon usage and occurs in over 90% of human genes [127, 128]. RNA splicing has been shown to be hugely adaptive: gene isoform ratios and splicing factor expression levels have been shown to change with age and have a transcriptional association with DNA damage repair factors in later life [129], while pan-cancer analyses have shown tumour cells to exhibit up to 30% more alternative splicing events than normal samples with a substantial portion unique to cancer subtypes [130].

In our blood, several studies have shown that specific repertoires of differential splicing are found in progenitor populations across haematopoietic lineages, suggesting that spliceosomal regulation of gene expression and associated proteomic changes may be required to define haematopoietic cell identity [131–133]. Mutations in splicing genes are common in many myeloproliferative disorders and occur in around 50% of patients with MDS [134], in particular the aforementioned *SF3B1*, *SRSF2* and *USAF1* are thought to occur in the early stages of the disease [135]. Mutations in these genes tend to cluster within specific functional domains and amino acid positions suggesting a tendency towards gain-of-function events and occur with a high degree of mutual exclusivity; indicating that mutations in these genes may involve some redundancy or exhibit a low systemic tolerance [136, 137]. One might assume this could indicate a shared pathogenesis, however, different mutated spliceosomal genes exhibit distinct

mechanistic characteristics as well as detectable changes in cellular morphology in MDS subtypes [138].

The precise workings of their role in driving clonal expansion remains elusive and can be assumed to be complex given their role regulating the transcriptome. Recent work has pointed to the cryptic splicing of differentiation and cell-cycle genes that provide a fitness advantage for *SF3B1* mutant cells within the erythroid lineage [139]. Perhaps the most interesting facet of spliceosomal mutations in the context of CH is their extreme age-dependence. Across a range of key CH population studies, spliceosomal mutations were observed uniquely above the age of 70 in low numbers but with relatively high VAFs using exome-sequencing [1, 2, 51], with more sensitive sequencing methods finding greater numbers above 70 at lower VAFs [52]. This is concordant with the sharp rise in MDS incidence in late life that is driven by mutations in the spliceosome [140] and that, within MDS patient populations, carriers of splicing variants tend to be significantly older [141]. Due to the stochastic nature of mutagenic processes, it is unlikely that the origin of these mutations occurs exclusively in later life and more probable that cell-extrinsic age-dependent changes precipitate and accelerate their growth in old age.

1.2.3.5 DNA Damage Response

Tumour Protein 53 (TP53)

Somatic variation in the TP53 gene is common in clonal haematopoiesis. TP53 is a potent tumour suppressor which responds to a host of cellular stressors and can facilitate a diverse array of effector pathways that protect genome stability and cell homeostasis [142]. The activation of TP53 is hugely context dependent and can transcriptionally regulate many hundreds of genes across several biological processes, including DNA damage repair (DDR) [143], senescence [144] and apoptosis [145].

In haematopoiesis, TP53 loss has been shown to provide a fitness advantage to both young [146] and old HSPCs [147] in a dose-dependent fashion. Mutations in DDR

genes (including *PPM1D*) are predominantly associated with patients who have undergone cancer treatment [55, 148] and now display clonal outgrowth, linking cytotoxic stress with clonal selection through increased resistance to apoptosis and a preference to cell-cycle activation [149]. Outside this context, DDR related mutations have been shown to be relatively common, including in young individuals without prior links to cancer therapies, suggesting that conferral of any fitness advantage might not only be linked to resistance to DNA damage or perhaps that cell-intrinsic levels of DNA damage can exert a positive fitness effect.

1.2.3.6 Additional Driver Event Classes

While we have examined the backgrounds of some of the most frequently mutated genes and gene classes in CH, there is an additional (although not insignificant) number of variants that have been discovered in population studies that can be considered putative drivers of CH.

Firstly, cohesin complex proteins, including RAD2, SMC3 and STAG2 and associated insulator CTCF have been implicated in both clonal haematopoiesis and AML [150, 151] and together, form a multiprotein complex involved in maintaining and manipulating 3D genome organisation, regulating gene transcription [152]. This gene class has been shown to be essential in maintaining the HSC population and the genomic architecture linking regulatory loci, with mutations affecting lineage specific differentiation [153]. Mutations in developmental transcription factors, like *RUNX1*, *CUX1*, *GATA2* and *NOTCH1*, likely act through similar mechanisms - disrupting the regulation of normal haematopoiesis [154].

Recent work using a sophisticated colony barcoding technique to determine clonal phylogenies discovered many large clones without known driver mutations [155] and the future discovery of novel CH variants will likely require significantly larger cohorts [156]. Many of these putative CH drivers may exhibit relatively weak fitness advantages and may require some polygenic co-operation to achieve substantive growth effects.

The acquisition of large mosaic chromosomal alterations (mCAs) has emerged as a hugely prevalent form of clonal mosaicism. The haematopoietic loss of the Y chromosome in men (mLOY) has been associated with CHIP and has a similar age-dependence – detectable in 40% of men by the age of 70 then rising to nearly 60% by the 10th decade of life – with correspondingly poor outcomes [157, 158]. Significant numbers of smaller mCAs have been observed in autosomal chromosomes and have been shown to have some interaction with haematological cancers [159] and a causal association with smoking [160].

1.2.3.7 Germline Determinants of CH

With the arrival of nation level genomic studies, the inherited factors that shape the growth of clonal haematopoiesis are beginning to be untangled. Two independent analyses of the UK Bio Bank, totalling over 200,000 and 600,000 participants respectively, have identified several novel heritable loci that may impact the growth of CHIP [156, 161].

Interestingly, the strongest germline association across all CHIP subtypes was shown at the *TERT* locus, a catalytic subunit of the telomerase enzyme involved in the maintenance and elongation of telomeres [162]. Telomere length is inversely correlated with age across tissues due to the gradual attrition of telomeres over repeated cycles of cell replication and gene expression [163]. Inherited SNPs in this gene tend to associate with maintenance of longer telomeres which may contribute to an increased resistance to mutation (via reduced propensity to senescence and terminal cell-cycle exit) and likely allows for increased rates of cell cycling [164].

Many other inherited risk loci overlap with genes that involve cell-cycle regulation and DNA damage response – with several displaying overlapping functions within these axes [165]. Germline risk in CHIP typically allows for the extension of replicative limits or resistance to mutagenesis, which can then cooperate with acquired genetic lesions to facilitate somatic evolution in haematopoietic cells.

Some germline SNPs have been shown to substantially limit predisposition to CH. A SNP in the *TCL1A* gene promoter has been shown to significantly reduce the risk of *TET2*, *ASXL1*, *SRSF2*, *SF3B1* and *JAK2* driven clonal expansions (although not *DNMT3A*-mutant CH). TCL1A normally exhibits low levels of expression in wild-type HSCs, but becomes activated in the presence of *TET2* or *ASXL1* mutations via increased promoter accessibility - the rs2887399 SNP in the *TCL1A* promoter likely blocks this activation. Carriers of this variant show an 80% reduction in CHIP susceptibility, likely indicating that this gene plays an important role in clone growth and could provide a target for the development of therapies to treat/prevent CHIP in a diverse genetic background [166].

1.2.4 Environmental and Extrinsic Drivers of Clonal Haematopoiesis

We have described how almost all individuals will harbour mutation driven HSC clones by late adulthood, however, only in a smaller proportion will the expansion of these clones be detectable, and ergo, potentially pathological. Alongside chance and time, exogenous factors play a significant contribution to clone growth as well as providing an explanation for the marked variation in prevalence and clone size seen amongst individuals [167]. With the arrival of a new set of environs, mutations that were once neutral now exhibit a fitness advantage, suggesting that some variants might only develop with a particular set of extrinsic signals (Figure 1.5).

The spatial structure and cellular composition of the bone marrow niche exhibits marked changes as we age, with these multifaceted changes potentially facilitating clone growth. Several studies have identified changes that include changes in the stromal cell compositions, a decrease in the bone matrix, increases in vascular volumes and expansions of large adipocytes that compound to create a space and environment markedly different from young bone [168–170]. How HSCs localise to other cell types has been known to change with ageing [171], with the decline in growth factor concentrations - like age-dependent reductions in IGF1 signalling across middle and into old age – now thought to contribute to a loss of HSC health-span and function [37].

Some of the best evidence for the importance of cell-extrinsic factors comes from observations in allogenic transplantation procedures in siblings [172, 173]. Some studies have shown that clonal haematopoiesis originating in the donor has been identified in the patient receiving the transplant and, in several cases, donor clones have shown marked enlargement in the reciprocant [174]. This perhaps suggests that either the transplantation or novel interactions with reciprocant bone marrow milieu promotes increased clonal outgrowth in these individuals. While these initial studies were limited to handfuls of patient pairs, larger studies in recent years have now highlighted a range of clinical outcomes that result from donor-CH are dependent on the mutational context and might require additional screening in the future [175].

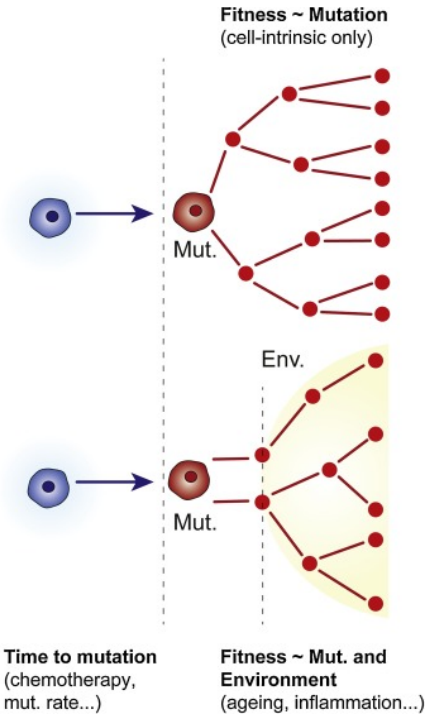


Figure 1.5: Is CHIP dependent on the environment or driven by cell-intrinsic factors? ***Top panel: CHIP is driven in a cell-intrinsic manner.*** *Here, time to acquiring the CHIP mutation (Mut.) and the subsequent change in selective advantage conferred by the variant are the key factors in clonal expansion. Average time to the mutation is dependent on a number of factors, including sequence context of the mutation, mutation rate and genotoxic exposures, such as chemotherapy.* ***Bottom Panel: Clonal expansions are potentiated by extrinsic factors.*** *In this model, the time to CHIP mutant (Mut.) acquisition and subsequent fitness are enabled by extrinsic factors (yellow background), such as exposure to inflammation, infection or age-dependent changes in the bone marrow niche.*

Chemotherapies and other cytotoxic treatments may provide a positive fitness advantage to mutant CH clones with rates of CH some 5-10 times above the levels observed in matched control groups [55, 176]. Patients who have undergone exposure to genotoxic agents display a particular enrichment for clones baring PPMD1, TP53 and other DDR related genes. Genotoxic therapies effectively destroy cancerous cells via the induction of DNA damage or impairing DNA damage response: it's likely that these variants in DDR genes enable cells to develop a growth advantage through resistance to this new mutagenic environment [149, 176].

Recent interest has surrounded the role of inflammation which - in many tissues - can become a chronic feature of old age. HSCs respond to a range of cytokine or chemokine triggers that can signal a requirement to replenish the blood system. The function of normal (wild-type) HSCs is detrimentally affected by the exposure and subsequent response to low-grade chronic inflammatory signalling that might gradually deplete or impair the HSC pool upon prolonged exposure [177]. Age-related exposure to chronic interleukin-1B (IL-1B), interleukin-6 (IL-6) or tumour necrosis factor alpha (TNF-α) can induce lineage biases, promote survival and exhaust stem cell function [178, 179]. TET2-mutant HSCs are refractory to inflammatory signalling and retain their functional integrity, thus providing a fitness advantage [180]. Furthermore, TET2 is also known to down-regulate inflammation in myeloid cells through suppression of IL-6, further potentiating the fitness advantage of TET2-mutant clones through the creation of an inflammatory feedback loop [94] with carriers exhibiting increased levels of circulating cytokines [181]. Similarly, inflammation related to acute and chronic infection can drive the expansion of DNMT3A-mutant clones via induction of IFNγ [85, 182].

There's an increasing body of evidence that suggests that there is an interplay between endogenous and exogenous drivers of ageing and clonal haematopoiesis, often in the form of inflammatory crosstalk between HSCs and their niche [183]. Ho et al., looked at a number of rejuvenation methods that are assumed to promote longevity, including parabiosis experiments. He found that while many of these methods can improve systemic function through the nutrient sensing axis, the blood system is refractory to these benefits, and interestingly, that young HSCs apparently

display minimal phenotypic changes when exposed to systemic aged environments, suggesting that young HSCs possess some intrinsic tolerance to inflammation [184].

1.2.5 Clonal Haematopoiesis and Disease Risk

An important focus of research in clonal haematopoiesis has been its links with myeloproliferative disease as a result of the enrichment of somatic mutations in many key driver genes. Recent focus has also surrounded the association to many non-haematological diseases that include many distal pathologies of ageing.

1.2.5.1 Associations with Haematological Disease

Most haematological cancers are characterised by the presence of cytopenia's – a reduction in specific peripheral blood counts – which are absent in CH carriers and may never be acquired [185]. Haematological malignancies such as AML and MDS can be seen in a similar context to many solid tumour types, which require the step-wise acquisition of several variants to achieve full oncogenic potential. Exactly how clonal haematopoiesis progresses through myelodysplastic syndromes, to AML and CML is poorly understood, however, it is thought that these "founder" mutations provide a fitness or proliferative advantage which increases clonal size and decreases clonal complexity, while impairing differentiation. Ensuing mutations hijack the fitness advantage of CH, then alter the function and output of the haematopoietic system terminally reducing the supply of mature blood cells which results in immunodeficiencies through decreasing lymphocytosis (CLL) or malignancies within the myeloid lineage (AML) [186].

The ordering of mutations is an important facet in most cancers and can be visualised as a form of branching evolution as the functional consequences of each variant may vary [187]. Second or third hits might be considered excessively deleterious in isolation and trigger senescence or apoptosis leading to the rapid demise of a potential clone – the founding mutation creates the intrinsic or extrinsic contexts that allow it to thrive [188]. Large studies of AML patients have indicated a substantial burden of DNMT3A

mutations that are typically associated with secondary hits in FLT3 (receptor tyrosine kinase) alongside mutations in NPM1 and RAS pathway kinases (NRAS, KRAS), suggesting that substantial cooperation is required to achieve pathogenicity [82]. Conversely, mutations in the spliceosome, in particular SF3B1, are thought to be the singular driver in over 30% of MDS cases [140]. JAK2-V617F mutations are observed in isolation in several MPN classifications, however the presence of cooperating mutations leads to substantially worse outcomes [189].

One might assume that due to the near ubiquitous nature of CH, cancer incidence would be low thus reflecting its high population prevalence. However, population studies have garnered that CH associates with an approximate 10-fold increase in risk for malignancy, depending on the sensitivity of the study [2, 190], with the mutational context [53] and the size of the driver clone a substantial contributor to this risk [3]. The prevalence of CH and risk of oncogenic transformation highlights the potential importance of population monitoring for this pre-malignant state.

1.2.5.2 Associated Risk with Non-Haematological Disorders

Clonal haematopoiesis is relatively unique in its capacity to cause chronic disfunction in other distal systems as HSCs and their progeny that are carriers CH driver mutations; a) lack the spatial constraints of solid tissues [191] and can expand more readily; b) constitute the main axis of adaptive immunity and function in every tissue with systemic outcomes, and; c) generally have higher inflammatory outputs [192].

CH is associated with increased mortality, not exclusively delineated through its links to blood cancers, but also via links to other distal pathologies [1]. When excluding the risk posed by haematological disease, CH (VAF > 2%; or 4% of circulating blood) can account for a 40% increase in excess mortality that has been shown to associate with cardiovascular disease (CVD) and ischaemic stroke [192] and has been recapitulated in a follow-up case-controlled cohort [97]. It has also been shown that increased clone size exacerbates these effects: patients with large clones (VAF > 10%) had a 12-fold heightened risk of coronary artery disease compared to non-CH participants [1] with the conferred risk posed by CH even comparing to well-known CVD risk factors, such

as hypertension, smoking and high cholesterol levels. While the exact causal mechanisms remain to be elucidated, two groups have assessed the interaction between CVD and CH using a murine model of accelerated atherosclerosis development – with a complete knock out of the low density lipoprotein receptor (Ldlr-/-) - and demonstrated that Tet2-mutant mice develop substantially larger plaques [97, 98]. Gene expression analysis showed that loss of Tet2 upregulates a host of inflammatory outputs, including Il1b, Il6 and members of the Cxcl family of inflammatory chemokines in circulating macrophages that integrate at the site of the lesion [98]. Subsequent treatment with small molecule inhibitors for Il1b and Il6 greatly reduced sclerotic plaque size in Tet2-mutant mice over their wild-type counterparts, suggesting a causal link with CH driven inflammatory outputs [98].

Furthermore, recent studies have suggested that clonal haematopoiesis is associated with congestive heart failure [193, 194], with Jak2, Tet2 and Dnmt3a mutant mouse models exacerbating cardiac failure through similar forms of inflammatory crosstalk [123, 195, 196]. Genetic studies in the UK Bio Bank have also begun to unravel the association between CH and atherosclerosis in humans. A well described loss-of-function polymorphism in the IL6R gene has been shown to confer a resistance to inflammation in DNMT3A- and TET-mutated CHIP and exhibit reduced CVD and mortality risk in individuals with large clones [197, 198].

Several relationships with CH and other pathologies have been described - including to COPD and diabetes – but these findings have proven difficult to replicate [199]. Perhaps the emergence of clonal haematopoiesis in late life allows it to lend itself as an able marker for biological ageing, and thus, provides a myriad of confounding features that are challenging to account for in association studies. An important question going forward will concern how CH mutations can affect distal tissues as carriers of complex phenotypes, driving pathology by perturbing the normal homeostasis of inflammation and function. It also highlights that, beyond the traditional risk of cancer progression, CH has broad effects on health and disease and that interventions to stabilise clone size or reduce their inflammatory output may eventually become a common therapeutic option depending on the age, medical context and preferences of the individuals involved [200].

1.2.6 Deciphering the Growth Potential of Mutations

The study of CH in large cross-sectional cohorts has provided a wealth of perspective on the genetic drivers, prevalence and associations with numerous clinical features. However, cross-sectional studies – providing a single snapshot in time across a population – leave numerous questions regarding how CH develops, the dynamics of clone growth and how it might interact with ageing. The earliest proxy used to understand these dynamics was a simple assessment of clone size, however such estimates erroneously assume similar points of origin and that the fitness effects conferred by mutations are similar. Therefore, new and more sensitive methods are required to determine accurate fitness estimates of clone growth.

The first longitudinal study in CH was relatively limited, containing a handful of individuals and two time-points and noted only the apparent heterogeneity between the growth of clones between individuals [64]. The first true quantitative analysis was performed by van Zeventer and colleagues, using a highly sensitive error-corrected targeted sequencing assay. This group was the first to quantify the longitudinal differences in VAF between clones in the different mutational contexts and noted that in *TP53*, *ASXL1* and *TET2* genes, there were significant differences in the change in absolute VAF [201].

The first truly systemic study into gene-specific fitness dynamics occurred in 2019. Watson et al. aggregated numerous large cohorts totalling over 50,000 blood samples from cancer-free participants. They leveraged the scale of this cohort, to calculate the fitness advantage of mutations by modelling the clone size distributions of the variant allele frequencies (VAFs). They combine several novel components in their algorithm: integrating mathematical models that concern the evolution of probability density functions under conditions of growth or drift with classic models of stochastic birth death processes to infer clone fitness at gene or even specific variant contexts provided there is sufficient mutational coverage [202]. Due to the cross-sectional nature of the data employed, the inferred fitness effects have relatively high error rates and it is difficult to untangle the possibility of competing or cooperating mutations.

However, this work provided the first comprehensive catalogue of fitness estimates and delineated the power of positive selection within the haematopoietic system.

Watson et al. made another interesting observation concerning the fitness distributions of synonymous variants in their model. Synonymous mutations confer no changes to the translated protein sequence and would normally be assumed to have a neutral or negative fitness advantage. In their data, many of the detectable synonymous mutations exhibited growth rates beyond what's expected of natural drift indicating that they must be passengers on clones with unknown driver mutations. A subsequent manuscript from this group has used the fitness effects of synonymous passengers and shown that genes with smaller fitness effects likely occur early, while rapidly growing high fitness variants tend to emerge in later life [203]. This would appear to have significant consequences for our understanding of CH, linking positive selection and cell extrinsic age-dependent effects.

In the last year, a competing paper to the work I present below was published, looking at a large multi-timepoint longitudinal cohort of CH. While deploying a different methodology, they similarly quantify the fitness effects of observed genes and come to similar conclusions. They highlight the rapid growth of spliceosomal genes in late life alongside the lifelong slow growth of DNMT3A driven clones. While our cohort tiers towards late life thus capturing CH at its most prevalent, their cohort tilts younger, representing a greater breadth of human lifespan. From this, they observe that fitness estimates are not stable across the life-time of a clone: some slow growing clones might require increased rates of growth in early life to reach their observed sizes [101].

1.3 DNA Methylation and Epigenetic Clocks

1.3.1 The Role of DNA Methylation with Age

Across the length of the mammalian genome, there are around 28 million unevenly distributed CpG dinucleotides (5'—C—phosphate—G—3'). In most of the non-coding genome CpGs are depleted, however, they are often enriched in clusters around gene promoters in what are known as CpG islands. CpG dinucleotides are the focus of one of the most abundant epigenetic modifications: the addition of a methyl group (CH_3) to create 5-methylcytosine (5mC) – also known as DNA methylation [204]. DNA methylation (DNAm) has a host of divergent functions that can include the maintenance of genomic stability through repression of transposable elements [205], regulation of transcriptional elongation and RNA splicing [206] and primarily the regulation of gene expression patterns [207].

DNA methylation is a highly dynamic modification. 5mC patterns can be constituted and eliminated by the DNA methyltransferase (DNMT) and ten-eleven translocation (TET) families of enzymes, respectively [208]. DNA methylation states can persist across cell divisions through the function of DNMT1, which copies methylation patterns onto the newly synthesized complementary strand [208]. While de novo methyltransferases, DNMT3A, DNMT3B and DNMT3L, catalyse methylation in non-replicating conditions important for differentiation and developmental programs [209]. TET enzymes passively remove 5mC via hydroxylation of 5mC to 5-hydroxymethylcytosine (5hmC) [87], while 5mC can additionally be eliminated by ineffective copying of the hemimethylated strand in DNA replication. Within post-differentiated tissues comprising many non-dividing cells such as the brain, liver and lung, DNAm is thought to be hugely dynamic. In such tissues, DNA methylation levels have been shown to actively oscillate at enhancers and promoters, altering transcription factor accessibility and effecting the dynamics of gene expression programs [210, 211].

Age-associated changes in DNA methylation have been described for many years and can occur through stochastic and active mechanisms [212]. Passive mechanisms

involve the ineffective maintenance of hemimethylated CpGs upon mitosis or poor provision of the metabolites that are needed to maintain consistent DNAm patterns [213]. These random and age-dependent increases (hypermethylation) or decreases (hypomethylation) in DNAm are known as epigenetic drift [214]. This drift creates its own form of somatic mosaicism within aged tissues, inexorably altering the fidelity of gene expression, reducing cellular plasticity and tissue function. Epigenetic drift has perhaps been best described through experiments that have focused on monozygotic twins. While DNAm levels track closely between the siblings through the early stages of life, old age leads to the development of substantial variance in DNAm levels – with particularly marked changes at CpG islands [215].

Not all age-associated changes in DNAm appear to be stochastic. It has been shown that with age and across tissues, similar patterns of hypermethylation can occur at a subset of stem-like Polycomb group target genes with similar signatures seen in some cancer subtypes [216]. Additionally, studies in mice have shown that enhancer regions significantly hypomethylate with age and that these changes can be attenuated with interventions that are known to improve longevity, such as calorie restriction and rapamycin treatment [217, 218]. The overlap between time-dependent DNAm changes and specific functional loci suggests that epigenetic changes play an important role in the deterioration of cell function with age [219].

1.3.2 DNA Methylation as a Predictor of Age

Ageing is the primary risk factor for numerous diseases, including cancer, cardiovascular disease and stroke. Therefore, it is imperative to understand the multitude of complex causal factors that inexorably lead to the physiological decline we see in old age. It is also clear that we do not all age at the same rates, and therefore, we need accurate biomarkers that allow us to accurate calculate the rate of biological ageing in individuals [204, 220].

In recent years, DNAm has emerged as a highly effective biomarker for predicting biological age. In the last decade, numerous DNAm age predictors, also known as epigenetic clocks, have been developed which utilize age-dependent shifts in DNA

methylation at single CpG sites to estimate biological age [221, 222]. Penalized regression models such as LASSO or elastic net are commonly used to select CpGs that have linear relationships with age in a given training dataset [223]. These selected CpGs are weighted and used to create an equation to estimate chronological age based on the percentage DNAm at “clock” CpG sites. Epigenetic clocks can capture different aspects of the ageing process and their increasing diversity have enabled quantitative methods of studying ageing [204]. For instance, some composite epigenetic clocks can estimate not only chronological age but also time-to-death, which can help predict morbidity and mortality [222].

Epigenetic clocks have proven to be adept at predicting the age of an individual, sometimes described as their chronological age (chAge). However, in some of the earliest clocks it was observed that the predicted epigenetic age (eAge) substantially deviated from the chronological age in some individuals [224, 225]. This posed a fundamental question: is this difference between the measured and predicted ages driven by inaccuracies within the model, or is it caused by biological factors (for example: disease burden, lifestyle or genetics) and therefore a biological clock?

While still relatively poorly defined, biological age intends to capture the functional decline of an organism through the prism of disease, morbidity and even mortality [221, 226]. Therefore, individuals who have identical chronological ages might have vastly different biological ages as a result of divergent health, lifestyle or disease profiles. Because of this, biological age is an important concept, as it provides a window with which we might be able to predict and assess disease risk and reflect both the qualitative and quantitative aspects of the ageing process [204].

It has been proposed that epigenetic clocks capture, at least in part, some aspects of biological age [220, 222, 227–229] and there is a growing body of evidence to support this. Accelerated eAge has been significantly associated with numerous syndromes and pathologies that have substantial functional and mechanistic evidence of advanced biological ageing, including HIV [230], Down syndrome [231] and Werner’s syndrome [232]. Conversely, long-lived individuals such as super-centenarians have been shown to exhibit a decelerated DNAm clock, indicating a lower biological age [233]. Many pathologies normally associated with advanced age have been shown to

associate with accelerated DNAm clock rates in cross-sectional studies. These include Alzheimer's and other neuropathies [229, 234], Parkinson's disease [235], non-alcoholic steatohepatitis (NASH) [236] even decreased physical and mental acuity in the Lothian Birth Cohort [237] to name but a few. Moreover, DNAm age has been used as a biomarker to predict the future risk of developing a range of disease states, including cardiovascular disease and cancer [228, 238–240]. The breadth of associations between DNAm clocks and a variety of age-dependent pathologies highlights the potential of epigenetic age as a candidate metric of biological ageing [218].

1.3.3 Epigenetic Clocks

The earliest epigenetic clocks developed were trained on relatively small numbers of samples and featured few CpGs. Blockand et al. trained the first clock on DNA methylation data derived from saliva in 34 pairs of twins and achieved a 5.2 year error rate in their initial age estimations [241]. Thereafter, Wolfgang Wagner's group then developed two methylation clocks - trained on multiple cell types and then on fibroblast culture passages - with limited accuracy [242, 243]. The first truly effective clock was developed by Horvath in 2013 which applied many of the commonly accepted design considerations used today [224]. Dubbed the first multi-tissue clock, it was trained across a large dataset of 8,000 samples from numerous sources and cell types, achieving an average accuracy of 3.6 years. The Hannum clock was devised with a similar strategy using peripheral blood and predicts age to an accuracy of 3.9 years with 71 CpG sites [225].

Epigenetic clocks have been typically trained on methylation data using penalized regression models. Such clocks can be represented as a linear model (Equation 1). They are predominantly built with Ridge, LASSO or ElasticNet regression algorithms which automatically select a set of key CpG loci – the "clock" sites – and assign a weight to each loci [244, 245]. These parameters are learned in penalised regression by minimizing the cost function and scaling the weights of uninformative CpG loci to 0 (Equation 1.1).

$$eAge \sim Intercept + \beta_1 CpG_1 + \beta_2 CpG_2 + \cdots + \beta_n CpG_n \quad \text{(Equation 1.1)}$$

The learned weights (β_n) for each site crudely describe whether a given CpG will monotonically increase or decrease with age for $\beta > 0$ or $\beta < 0$, respectively. Such penalized regression models reduce over-fitting where the number of available data-points used for training is large, such as in a 450k methylation array [204, 222].

DNAm clocks have proven successful at predicting age estimates when trained against the chronological ages of individuals. However, to capture differing aspects of the ageing process, clocks might need adaptations to their training to make them sensitive to a new set of phenotypic traits. These clocks have been termed composite clocks as they typically utilise a variety of additional biomarker data beyond a simple regression against chronological age. The first composite clock from Morgan Levine joined ten biometrics that displayed a significant association with age, including proteomic data (C-reactive protein, glycated haemoglobin, serum creatinine, urea and albumin) and phenotypic measurements (forced expiratory volume, systolic blood pressure) amongst others [246]. This was followed up several years later, when Levine et al. combined several physiological and phenotypic measurements with chronological age to generate the PhenoAge clock [229]. The 513 CpG sites that compose this metric have proven to be a more effective predictor of all-cause mortality, physical acuity, cancer and morbidity than any previously devised [229].

Similar strategies were deployed in the development of the GrimAge clock. Here the authors generate a set of clock CpGs by training against smoking status (in pack-years) coupled with seven blood serum proteins associated with shortened lifespan. The resultant clock proved to be effective at predicting mortality, cancer and coronary heart disease [247].

Due to the ease of accessing biological material, many clocks have been trained on peripheral blood mononuclear cells (PBMCs). But as described in Section 1.2, the haematopoietic system experiences significant changes with age that cause lineage skewing and may alter blood cell count proportions. To counter this, two methods have been devised that take different approaches to this issue: firstly, intrinsic age acceleration (IEAA) takes into account changes in blood cell count proportions and adjusts the measurement accordingly to remove them as a confounding factor in age

estimates; while extrinsic age acceleration (EEAA) takes the opposite approach by attempting to incorporate cell count proportions into age estimates [248, 249].

One final key element that substantially effects clock accuracy is the size and scope of the training dataset. The ZhangAge clock utilises 13,402 blood samples (as well as 259 from saliva) to construct one of the largest training datasets to date [250]. They achieve a remarkable correlation between chAge and eAge, achieving an accuracy of 2.04 years in their prediction. This highlights the importance of data curation: to achieve the most accurate predictions, you need a large number of samples with a sufficient age range and resolution to achieve accurate measurements of eAge across the breadth of the human life-span.

Clock	CpGs	Tissue	Training Size	Method	Type	Reference
IEAA (Horvath)	353	Multi-Tissue	8,000	ElasticNet	Intrinsic	Horvath (2013)
EEAA (Hannum)	71	Blood	482	ElasticNet	Extrinsic	Hannum et al. (2013)
PhenoAge	513	Blood	9,926	ElasticNet	Composite	Levine et al. (2018)
GrimAge	1,113	Blood	1,731	ElasticNet	Composite	Lu et al. (2018)
Zhang Clock	319,607	Blood (Saliva)	13,661	ElasticNet	Intrinsic	Zhang et al. (2019)

Table 1.1: Summary of the main human DNAm methylation clocks used in this thesis. *Highlighting a variety of differences between them, including training sizes, covariables and number of clock CpG sites used in the prediction.*

This thesis uses several of these published clocks to assess for changes in biological age that are summarised in Table 1.1.

1.4 Thesis Aims

It is now understood that potentially oncogenic mutations emerge with regularity in the blood of ostensibly healthy individuals. As we age, these mutations drive the expansion of clones in our HSC pool that leads to an elevated risk of both haematological and non-haematological disease. Currently, we have a limited understanding of the dynamics of clonal haemopoiesis, how it cooperates with ageing and how these components might link to disease.

This thesis attempts to answer the following questions:

1. Clonal haematopoiesis presents in later life and associates with many distal pathologies linked to advanced age. We utilise several distinct DNA methylation clocks to ask: do carriers of clonal haematopoiesis exhibit accelerated biological ageing?

2. Clonal haematopoiesis is an expansion of a single HSC driven by positive selection in a functionally diverse set of genes. Can we untangle the fitness advantages of clone growth and its variation between individuals? Additionally, can we leverage fitness advantages as a new predictor of disease risk?

Chapter 2: Study Cohorts and Methodologies

In this chapter, I provide a broad summary of the protocols and design constraints used in the development of the main cohort used in this thesis: the Lothian Birth Cohorts of 1921 and 1936. Furthermore, I describe the methodology used to classify somatic mutations, the use of paired DNA methylation data and its application to DNA methylation clocks in these cohorts. Finally, I describe the set-up, selection criteria, processing and analysis of our longitudinal targeted sequencing panel of genes that commonly present in individuals with clonal haematopoiesis and the methods deployed to extract the fitness advantages (or growth speeds) of HSPC somatic clones.

2.1 The Lothian Birth Cohorts of 1921 and 1936

The Lothian Birth Cohorts (LBCs) of 1921 (LBC1921) and 1936 (LBC1936) are two parallel longitudinal studies of ageing and were primarily devised to study features of cognitive decline with respect to age. The participants of both cohorts consist of individuals born in 1921 and 1936 who undertook the Moray House Test No. 12 – a test of cognitive ability – at age 11. In all, Lothian Health Board recruited approximately 70,000 young people who were registered in the region. With time, this has provided a unique baseline of mental acuity in the population for when in 2006 surviving participants were recruited to form the body of what is now known as the Lothian Birth Cohort [251–253].

Beginning in the year 1999, surviving members of these original tests were written to and asked to participate in what was then to be the largest study of non-pathological cognitive ageing. In all, 550 participants were enrolled in the LBC1921 cohort (mean age of 79 years) and 1,091 in the LBC1936 cohort (mean age of 70 years).

The initial objectives for the LBC were to utilise these historic measurements to better understand the determinants of cognitive ageing. In this vein, similar cognitive testing to the original Moray House examination was performed, alongside a host of physical

examinations, life-style questionnaires and consolidation of medical, sociological and demographic records. Alongside this, efforts to store blood to assay for genetic and epigenetic associations were made [251]. Blood draws have been taken across five waves of data collection at the average ages of (LBC1936/LBC1921) 70/79, 73/82, 76/85, 79/88 and 82/91 years. The original study design is clarified in Deary et al. (2007) [251].

2.1.1 Ethics, Funding and Data Access for the LBCs

These studies comply with relevant ethical regulations. The study protocol was approved by NHS Lothian (formerly Lothian Health). Informed consent was given by all participants. Ethics permission for LBC1936 was obtained from the Multi-Centre Research Ethics Committee for Scotland (wave 1: MREC/01/0/56), the Lothian Research Ethics Committee (wave 1: LREC/2003/2/29) and the Scotland A Research Ethics Committee (waves 2, 3, 4 and 5: 07/MRE00/58). Ethics permission for LBC1921 was obtained from the Lothian Research Ethics Committee (wave 1: LREC/1998/4/183; wave 2: LREC/2003/7/23; wave 3: 1702/98/4/183) and the Scotland A Research Ethics Committee (waves 4 and 5: 10/MRE00/87).

Genome-wide DNA-sequencing was funded by the BBSRC (Biotechnology and Biological Sciences Research Council). DNA methylation arrays were funded by the Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, the Wellcome Institutional Strategic Support Fund, The University of Edinburgh and The University of Queensland.

LBC phenotypic data are available in the database of the Genomes and Phenomes (dbGAP) under accession number phs000821.v1.p1. All other Lothian Birth Cohort data are deposited in dbGAP or are provided via the LBC Data Access Collaboration. Information concerning the cohort is contained here, including its history, data summary tables for both LBC1921 and LBC1936 and data access request forms and contact information to obtain all data points.

2.2 Methodology to Assess the Association of Clonal Haematopoiesis and Accelerated Epigenetic Ageing

2.2.1 Selection from the Lothian Birth Cohorts

Participants were selected from the Lothian Birth Cohorts (LBCs). In all, 1,136 individuals have been included in this analysis that have paired whole-genome DNA-sequencing and DNA methylation (DNAm) data (Illumina Human Methylation 450K BeadChip). These participants are drawn from both LBC1921 (n=104 and n=166 at waves 1 and 4, mean ages 79 and 88, respectively) and LBC1936 (n=873 from wave 1, mean age of 70 years).

2.2.2 Calling Somatic Mutations in Whole Genome DNA Sequencing

The whole-genome DNA-sequencing was analysed as follows: Raw sequences were initially assessed with FastQC before filtering for poor quality reads (phred quality score ≥ 30) and read ends with Trimmomatic [254]. Sequence libraries were then aligned to the human reference genome GRCh38 (including alt, HLA and decoy sequence contigs) with the Burrows-Wheeler Aligner (BWA) [255]. Duplicate reads were removed with samblaster (v0.1.22) [256]. Cleaning and optimisation of the alignment files was conducted with the Genome Analysis Toolkit (GATK; v3.4.0) to optimise alignments around known “gold standard” SNPs and indels observed in the 1000 genomes project. Base recalibration was also performed to adjust for intrinsic biases in base phred quality scores that are introduced by sequencing platforms, ultimately permitting more accurate variant calling downstream [257]. This yielded a mean genomic coverage of 34.3 reads.

We called somatic variants and short indels with the MuTect2 suite (v3.8) [257, 258]. Candidate mutations were manually assessed for quality and over-represented variants – likely driven by sequencing errors or poor-quality alignments - were removed. Single nucleotide variants (SNVs) were then annotated with the COSMIC (Catalogue of Somatic Mutations in Cancer) database [259] using the Ensembl Variant

Effect Predictor [260]. VCFs were additionally transformed to MAF files (Mutation Annotation Format) using vcf2mf (v1.6.16).

To effectively characterise the presence of mutations known to be associated with clonal haematopoiesis, we compared our mutations to a list of driver mutations described in Jaiswal et al. (2014) and Genovese et al. (2014) [1, 2]. After curating against these lists, we identified somatic mutations in six key genes that exhibited the highest mutational burden or are most frequently associated with clonal haematopoiesis in the literature. This left 73 participants (6%) that we consider as exhibiting some form of mutation driven clonal haematopoiesis in the downstream analysis.

2.2.3 Processing and Normalisation of DNA Methylation Data

DNA methylation was profiled across 485,512 CpG sites using the Illumina Human Methylation 450K BeadChip. Raw methylation bead intensities were read into the R language using the minfi package [261]. Normalisation was conducted using the "Noob" (Normal-exponential using out-of-band probes) method which utilizes a background subtraction technique alongside dye-bias normalisation [262]. In short, this method estimates background noise profiles from out-of-band probes which it then removes from samples. Normalised DNAm matrices were then submitted for epigenetic clock analyses.

2.2.4 Epigenetic Age Estimators

Several epigenetic age estimates were calculated for all selected participants. A range of clocks were selected that are assumed to capture different aspects of the ageing process. Firstly, we considered the Intrinsic Epigenetic Age Acceleration (IEAA) measure, an adapted version of the original Horvath clock [228] that regresses out the effects of changes to blood cell count proportions as well as the Extrinsic Epigenetic Age Acceleration (EEAA) metric – that utilizes the original Hannum clock algorithm and clock sites, then couples estimates of age with white blood cell compositions [225, 263]. Two composite clocks were included in this analysis, including PhenoAge and GrimAge methods [229, 247] alongside the ZhangAge method [250]. It must be noted

that the ZhangAge clock was partially trained on the Lothian Birth Cohorts and therefore, there may be some bias and overly optimistic predictions in participant age estimates. The majority of the estimates were obtained using the online clock calculator [https://dnamage.genetics.ucla.edu/home]. The ZhangAge clock was partially developed in my host institution by the Marioni group - an advisor and corresponding author of this manuscript. These clocks are discussed in some detail in Section 1.4.3 and Table 1.1.

2.2.5 Covariates and Regression Model

To assess for associations between CH and epigenetic age we perform a regression analysis using linear models from the 'lm' function in the 'stats' library and 'glm' from the MASS packages in R. We adjusted our model using age and sex as covariates alongside several immune cell proportions imputed from the methylation data (monocytes, CD4T, CD8T, Natural Killer and B cells). We also assessed the relationship between CH and blood cell count proportions using age and sex as covariates [229, 264].

2.3 Methodology to Characterise the Dynamics of Gene Specific Fitness Effects

2.3.1 Participant Selection and Characterisation

In our previous study (Robertson et al., 2019), we identified 73 participants with clonal haematopoiesis (CH) in the Lothian Birth Cohorts (LBCs) [4]. While there have been numerous studies of clonal haematopoiesis in cross-sectional cohorts that have advanced our understanding of the prevalence of CH and its links to disease, at the time of writing this study, there had been few large longitudinal cohorts that had assessed the dynamics of CH over time. To this end, we have utilized an error-corrected targeted sequencing approach on a panel of 75 genes (Table 2.1) that are known drivers of CH and other myeloid malignancies like AML, MPN and MDS. We then sequenced DNA from LBC participants over 3-year intervals that cover the eighth and nineth decades of life.

Target Genes in Sequencing Panel						
ABL1	CBLC	DNMT3A	IDH2	MYC	RAD21	STAG2
ANKRD26	CCND2	ETNK1	IKZF1	MYD88	RBBP6	STAT3
ASXL1	CDC25C	ETV6	JAK2	NF1	RPS14	TET2
ATRX	CDKN2A	EZH2	JAK3	NOTCH1	RUNX1	TP53
BCOR	CEBPA	FBXW7	KDM6A	NPM1	SETBP1	U2AF1
BCORL1	CSF3R	FLT3	KIT	NRAS	SF3B1	U2AF2
BRAF	CUX1	GATA1	KMT2A	PDGFRA	SH2B3	WT1
BTK	CXCR4	GATA2	KRAS	PHF6	SLC29A1	XPO1
CALR	DCK	GNAS	LUC7L2	PPM1D	SMC1A	ZRSR2
CBL	DDX41	HRAS	MAP2K1	PTEN	SMC3	
CBLB	DHX15	IDH1	MPL	PTPN11	SRSF2	

Table 2.1: Summary of genes included in the targeted sequencing panel.

Cohort Information					
Wave Number	1	2	3	4	5
Mean Age (years; LBC1936 / LBC1921)	70 / 79	73 / 82	76 / 85	79 / 88	82 / 91
Number of Samples	40	40	78	63	28
Gender (m/f)	21 / 19	21 / 19	40 / 38	31 / 32	14 / 14

Table 2.2: Cohort information across the waves of LBC data collection.

Using this approach, we sequenced the 73 participants that had previously been identified and added an additional 16 with previously unknown CH status that had the most available time-points. We have included 85 of the 89 participants in our study,

removing 4 participants that generally exhibited low quality criteria or library complexities across the time-course. In all, 248 samples have been included that were sequenced in seven batches that included two “Genome in a Bottle” control samples per batch (14 total) [265]. All participants included in the study have between 2 and 5 time points (Table 2.2)

2.3.2 Targeted Error Corrected Sequencing and Data Filtering

Libraries were prepared by Prof. Lee Murphy, Angie Fawkes and Louise MacGillivray in the Edinburgh Clinical Research Facility. DNA was extracted from EDTA whole blood using the Nucleon BACC3 kit (Sigma Aldrich, cat. Nb GERPN8512), 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, cat. Nb. AB0108, VariantPlex-HGC Myeloid Kit, for Illumina), 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 (Illumina). To inform reproducibility, we sequenced two “Genome in a Bottle” DNA samples in each batch of samples (DNA NA12878 Coriell Institute) to create a background model for variant errors and batch correction [5, 265].

Sequence libraries were filtered for phred ≥30 using Trimmomatic (v.0.27) before alignment to human reference genome hg19 using bwa-mem (v0.7.17) [255]. To improve the quality and accuracy of VAF calls, the platform uses a set of unique molecular identifiers that are ligated to each read before PCR amplification to enable accurate read deduplication. Within the target gene panel, somatic mutations are called using three independent variant callers: LoFreq (v2.1.0) [266], FreeBayes [267] and Vision (unpublished). Consensus is required from two of the three callers to be included downstream.

All somatic mutations at 2% VAF met a series of quality criteria: 1) the number of reads supporting the variant has to surpass a coverage threshold with supporting reads

exhibiting no directional biases (AO ≥ 5, UAO ≥ 3); 2) mutations are significantly underrepresented (exhibiting low population VAFs) in the gnomAD database (Genome Aggregation Database; p ≤ 0.05) [268]; 3) mutations are not obviously inherited from the germline (stable allelic frequencies across time-points at ~0.5 or ~1 VAF) which might have been overlooked by the gnomAD dataset as a result of the small geographical origin of the Lothian Birth Cohort participants; 4) do not exhibit excessive overrepresentation across the dataset. Over-represented events were typically frameshift duplication and deletions – we believe that the reads driving these events share some sequence homology to the targeted regions and are likely misaligned artefact from the capture method. We have also curated and cross-referenced our variants with both the Catalogue of Somatic Mutations in Cancer (COSMIC) database and with reported CH drivers in two large cross-sectional studies [1, 2, 259]. Lastly, for any variant that surpassed VAF≥2%, we have included any additional participant matched data points across all available time-points regardless of their VAF levels (Figure 2.1).

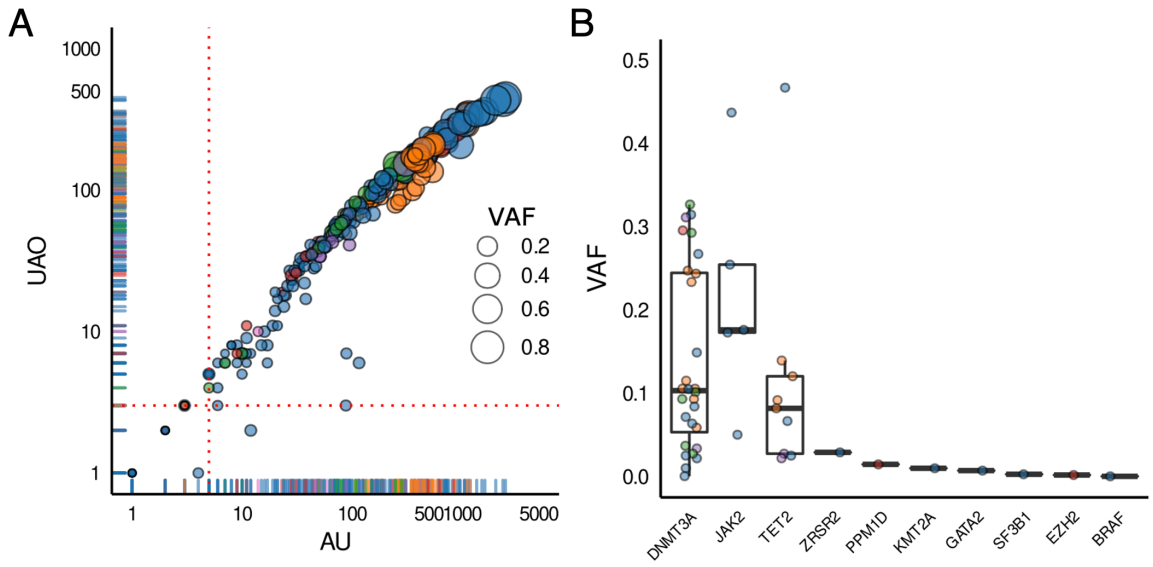


Figure 2.1: Quality control metrics: Coverage. ***A.*** *Sequence coverage metrics for variant calls across all participants and time-points filtered for 2% VAF. The displayed metrics are the AO (the number of reads that support the alternative allele, or mutation), versus the UAO (the number of sequenced reads with unique start sites supporting the alternative allele – a measure of molecular complexity). Red dotted lines indicate the filter thresholds for each measurement (AO ≥ 5, UAO ≥ 3). Points are scaled by the VAF of each call. Only 7 of 275*

variant calls failed to meet the required filter criteria, however, they were not excluded as they were supported with data matched events in any participants time-course. ***B.*** *Box plot showing median and interquartile ranges of the allele frequency of all observed mutations at Wave 1. This crudely represents how CH is captured in a traditional cross-sectional study.*

Sequencing artefacts can become highly problematic when attempting to detect variants at low VAFs in a targeted sequencing platform. To further reduce false-positive variants in our libraries we leverage the pan-dataset coverage levels of each sample and the "Genome in a Bottle" controls to generate a position-specific noise profile at each variant to accurately determine a limit of detection (LOD) for each variant. We report two additional parameters for each mutation: 1) the Minimal Detectable Allele Fraction (95% MDAF; Figure 2.2 A) which denotes the minimum VAF that a variant can be detected – essentially, describing a Limit of Detection (LOD) for each mutation, and; 2) the VAF Outlier P-Value which describes the probability that any mutation call might be an artefact driven by noise or error in the sequencing platform. This method calculates the position specific noise distribution across the Genome in a Bottle controls and across the pan-dataset coverage levels of all samples allowing us to discern over-represented sequencing artefacts from real events (Figure 2.2 B) [5]. All mutations that had matched time-series data that met our acceptance criteria were included downstream (VAF-Outlier P-Value ≤ 0.1).

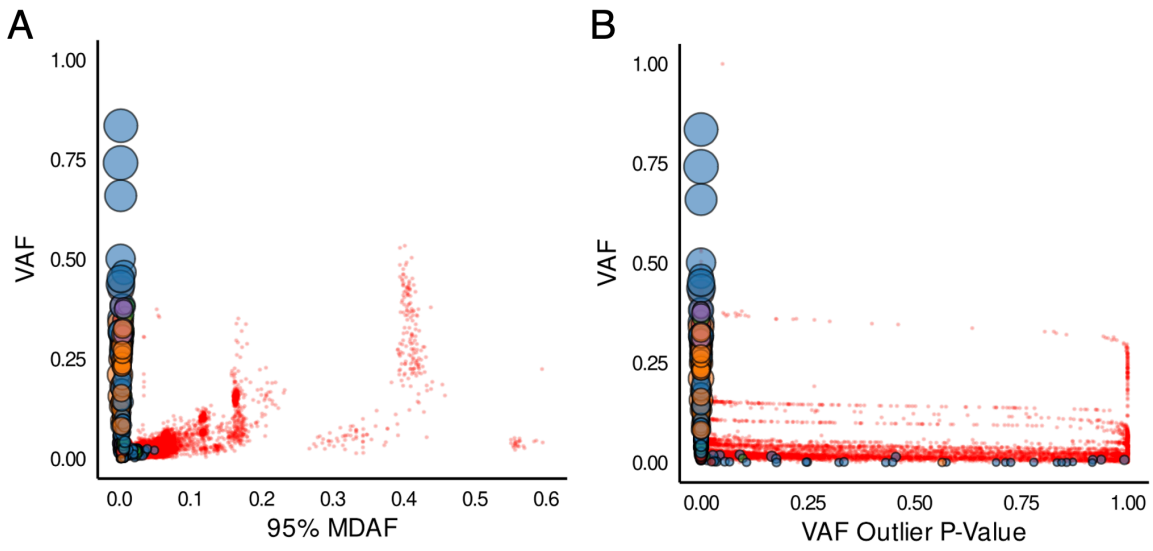


Figure 2.1: Quality control metrics: Error rates in captured variants. ***A.*** *The 95% MDAF (Minimal Detectable Allele Fraction with 95% Confidence) versus the VAF for each*

event. All variants used in our analysis above 2% VAF are scaled by their clone size and coloured by their functional consequence. Points in red are events that failed to pass our quality criteria and are removed from subsequent work. ***B.*** *The VAF Outlier P-Value (describing the pan-cohort position-specific background noise) versus VAF for each event. All variants used in the analysis above 2% VAF are scaled by their clone size and coloured by their functional consequence. Points in red are events that failed to pass our quality criteria and are removed from subsequent work. All accepted events that exceed VAF Outlier P-Value > 0.1 are generally low VAF and are supported by matching events across the time-series that adhere to our acceptance criteria of VAF Outlier P-Value ≤ 0.1.*

2.3.3 Computational Prediction of Missense Variant Effects

Predicting the pathophysiological consequences of mutations can be a non-trivial task. There are many gaps in our understanding of gene expression and its downstream effects and lack of well annotated data that link events to functional outputs. Most importantly, we still have an incomplete understanding of protein structure and function – which is particularly telling in some genes [269, 270]. To get an accurate picture of the effects of our variants, we have collaborated with Joe Marsh and Ben Livesey (University of Edinburgh). Their pipeline uses seven computational variant effect predictors recently identified as being most useful for identifying pathogenic mutations [271–277]. For each of the variants highlighted in this work, we calculated the fraction of previously identified pathogenic and likely pathogenic missense mutations from ClinVar and the proportion of variants observed in the human population in the gnomAD database (v2.1) for the outputs of each computational predictor before averaging across all predictions. The variant effect predictor Deep Sequence [276] was not run across all proteins due to the computational burden of running it against long protein sequences. Additionally, attempts to predict the effects of missense variant on protein (de)stabilisation were performed using FoldX (v5.0), using the experimentally determined protein structure (if available), or using the AlphaFold protein structure prediction [278, 279].

2.3.4 Mathematical Model of Clonal Dynamics to Infer Fitness

This part of the methodology was undertaken with substantial input from mathematicians and modellers Dr Eric Latorre-Crespo and Dr Linus Schumacher. Given the longitudinal nature of this study we can use the probabilistic solution of an established minimal model of cell division to infer the parameter distribution resulting in the observed time evolution of VAF trajectories in a participant's genetic profile [21, 202]. For each individual we simultaneously estimate the fitness of variants as well as the size of the stem cell pool, without needing to estimate the time of mutation acquisition. 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 parameter s and determines the fitness advantage of a clone. In normal haematopoiesis $s = 0$, in which case clones undergo neutral drift. For clones with non-neutral (fitness-increasing) mutations, $s > 0$, and these average clone size grows exponentially in time as $e^{s(t-t_0)}$ from an initial population of 1 SC at the time of mutation acquisition t_0. The full distribution of clone sizes is well approximated by a negative binomial distribution matching the mean (exponential growth) and variance of the full stochastic solution. Since the model dynamics are Markovian (without memory), once we condition on a previously observed time-point in a trajectory, the prediction for all future times is independent of t_0. From the predicted clone size distributions, we can infer the marginal posterior distribution of parameter *s* using

Bayes' theorem. We further take into account the sampling error during sequencing to estimate the distribution of clone sizes at the start and end of each time interval in the longitudinal sequencing data. Here we approximate this sampling error as binomial.

When multiple fit clones are present in an individual, we constrain the inference 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 the evolution of exponentially growing clones can be suppressed by the growth of a fitter clone. 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.

We take into account possible clonal substructures for all fit variants in an individual, selecting models with co-occurring mutations on the same clone if they are more likely after biassing against models with multiple mutations per clone, as these are presumed to be rarer. We then infer the posterior fitness distributions per clone for the most likely clonal model in every participant.

Once we have inferred the posterior distributions of the parameters, we use the mode of the distribution (maximum a posteriori (MAP) estimate) for each mutation to visualise the deterministic, i.e., average, growth curves. These result in the logistic time evolution of its corresponding VAF,

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

where we determine the time of mutation acquisition t_0 - which is only used for plotting - using maximum likelihood. Although deterministic fits are not a direct reflection of the inference results of our stochastic model, these can be used to visually assess the "goodness of fit" of the fitness MAP estimates and have been included for each participant in the LBC1921 and LBC1936 respectively in later figures.

Note that this model cannot account for loss of heterozygosity events.

2.3.5 Likelihood-Based Filter for Time-Series Data (LiFT)

To select fit variants, we compare the likelihood of the clonal model, including binomial sampling error, to a model of sequencing artefacts. The artefact model assumes all variability arises from sampling error with a proportion that remains constant over time. For variants that occur more than once in our dataset we use a beta-binomial model to account for overdispersion and for unique variants we use a binomial model. We select variants as fit only if the model evidence for the clonal model is at least 4 times that of the artefact model. Fit variants thus selected are taken through to clonal structure model selection and fitness inference as described above.

2.3.6 Framework and Data Availability

Both LiFT and Bayesian inference of the posterior distribution of model parameters were implemented in Python v.3.7 [280] with dependencies on NumPy v.1.21.5 [281], Scipy v.1.7.3 [282] and Pandas 1.3.4. Survival analysis was implemented using Python v.3.7 with dependencies on Lifelines 0.26.4. Data curation was undertaken in Python v.3.7 and R base.

All read data from the longitudinal cohort has been deidentified and uploaded to the NCBI Gene Expression Omnibus (Geo) with accession ID: GSE178936.

Chapter 3: Clonal Haematopoiesis is Associated with Accelerated Epigenetic Ageing

3.1 Introduction

The gradual accumulation of genetic damage is one of the hallmarks of ageing [283]. Clonal haematopoiesis (CH) is a characterised by the emergence of clonal subpopulations in our circulating blood and HSPCs. The expansion of these clones is driven by somatic mutations in genes that lead to increased clonal fitness in the stem cell population. While CH is defined by the presence of clones in apparently healthy individuals lacking obvious cytopenia's, the genetic drivers are also known drivers of myeloid malignancies [53, 185]. The prevalence of CH markedly expands with age to become almost ubiquitous by late-life. Alongside this, clonal haematopoiesis has been shown to associate not only with cancer, but with many distal pathologies, including ischaemic heart failure, cardiovascular disease and atherosclerosis for which age is a key risk factor [1, 194].

While CH has a strong association with several diseases associated with ageing, it's links to accelerated biological or epigenetic age remain unexplored. DNA methylation (DNAm) has emerged as a powerful tool for estimating biological age and numerous DNAm clocks have been developed that use a collection of clock CpG sites that track with age [223]. The vast majority of these clocks have been built with penalised regression algorithms that select these clock sites based on the linear relationships of DNAm changes with age, the resultant weighted algorithm across selected CpGs have proven to be adept at predicting the epigenetic age of test participants [204]. Deviations from chronological age towards an increased epigenetic age have been associated with increased risk of all-cause mortality and age-related morbidities [220, 222, 227–229]. Accelerated epigenetic age has also been shown to be associated with numerous pathologies that have functional and mechanistic evidence of advanced biological ageing, highlighting the capability of epigenetic clocks to effectively capture elements of biological ageing [230, 232, 234, 235].

Several epidemiological and functional studies have linked clonal haematopoiesis to distal pathologies of age, here we used paired whole-genome sequencing and DNA-methylation data from the Lothian Birth Cohort to present evidence of accelerated epigenetic ageing in individuals with CH.

3.2 Results

The Lothian Birth Cohorts of 1921 (LBC1921) and 1936 (LBC1936) are two longitudinal epidemiological studies of ageing [251–253]. Participants have been followed up every ~3 years, each for five waves, from the ages of 70 and 79, for the LBC1936 and LBC1921 respectively. Participants were community-dwelling, relatively healthy and mostly lived in the City of Edinburgh or its surrounding area when recruited.

Whole blood DNA methylation levels were assessed using the Illumina HumanMethylation450 BeadChip. Quality control details are reported in the Section 2.1.3. Genomic variants were determined in 1,136 LBC participants (n=870 from wave 1 at mean age 70 years in LBC1936; n=101 and n=165 at mean ages 79 and 87, respectively in LBC1921) where paired whole-genome sequencing (WGS) and methylation data was available. WGS data were aligned with Burrows-Wheeler Aligner and processed for duplicate mapping reads with samblaster yielding an average genomic coverage of 34.3 reads. Single-nucleotide variants and short indels were called with MuTect (v3.8) before annotation using the Ensembl Variant Effect Predictor alongside the Cosmic database of coding mutations (v86) [258–260]. CH variants were classified as per Jaiswal et al [1]. A detailed description of the methodology is included in the Section 2.1.

We considered and assessed epigenetic age acceleration in CH across six different DNAm clocks. Firstly, we utilized the Intrinsic Epigenetic Age Acceleration (IEAA – hereafter referred to as Horvath age acceleration) measure, which is an adapted version of the original Horvath clock that controls for white blood cell proportions [228] and has been considered to be partly driven by the number of cell divisions [284]. We have also assessed for accelerated epigenetic ageing using the adapted extrinsic Hannum clock (Extrinsic Epigenetic Age Acceleration – EEAA) which can track with

changes in blood cell count compositions [249] and can be considered to be influenced by external or environmental factors [229, 247]. Alongside this, we have additionally assessed for age acceleration with the Zhang Clock [250] and two composite clocks trained as predictors of all-cause mortality – the PhenoAge [229] and GrimAge models [247]. A detailed summary describing the training parameters and desired outcomes of the epigenetic clocks used in this analysis is shown in Table 1.1 and Table 3.1.

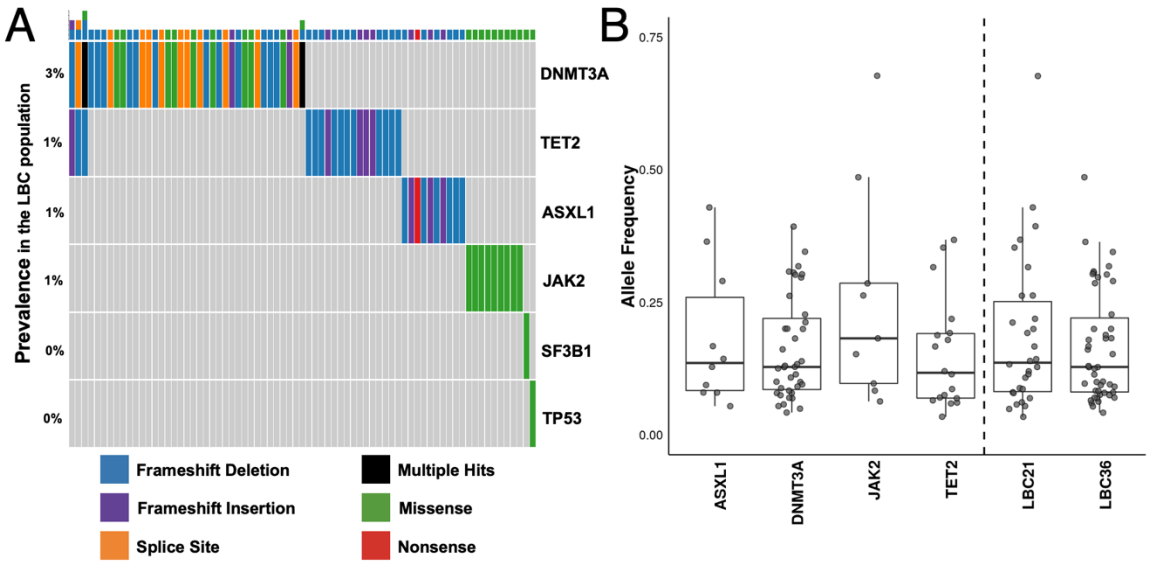


Figure 3.1: CH variants discovered in Lothian Birth Cohort (LBC) participants. ***A.*** *Oncoplot showing variant types within the CH positive subset of the LBC. This subset represents 73 participants (6% of 1,136 total) where one or more described somatic variants were detected in the six most prevalent CH-associated genes.* ***B.*** *Box plot describing the distribution of variant allele frequencies in all detected somatic CH variants. Genes with a single variant not shown are TP53 and SF3B1 (allele frequencies of 0.089 and 0.257, respectively). The overall distribution of allele frequencies by LBC cohort (LBC1921/LBC1936) is also shown.*

Epigenetic age estimates were regressed on chronological age to yield age acceleration residuals. Linear regression adjusting for sex, imputed white blood cell proportions (Monocytes, Natural Killer, CD4T, CD8T and B Cells) and methylation processing batch was used to determine the association between CH status (predictor) and Age Acceleration (response).

Clock	CpGs	Tissue	Outcome	Type	Training Details
IEAA (Horvath)	353	Multi-Tissue	Chronological age	Intrinsic	Primarily taken from non-tumour TCGA samples and reweighted to reduce effects of blood cell count proportions.
EEAA (Hannum)	71	Blood	Chronological age	Extrinsic	Utilizes same CpG sites as Hannum, but trained and reweighted to maximize influence of blood cell count proportions.
PhenoAge	513	Blood	Mortality	Composite (Extrinsic)	A measure of all-cause mortality, through training with several associated markers. These include blood cell proportions and morphology, C-Reactive Protein (CRP) and serum glucose, albumin and creatinine.
GrimAge	1,113	Blood	Mortality	Composite (Extrinsic)	This clock is trained on eight biomarkers that have been used to predict advanced ageing and mortality. These include, Adrenomedullin (ADM), Cystatin C, Leptin, SERPINE/PAI1, Growth Differentiation Factor 15 (GDF15), Beta-2-Microglobulin (B2M), TIMP Metalloproteinase Inhibitor 1 (TIMP1) and smoking pack years.
Zhang Clock	319,607	Blood (some Saliva)	Chronological age	Intrinsic	At the time of publication, this clock was trained on the largest available dataset and had the highest accuracy on internal test data.

Table 3.1: A summary of the training parameters and desired outcomes of the epigenetic clocks used in this analysis.

Of the ten most prevalent CH mutated genes described in several epidemiological studies (Figure 1.4) [1, 2, 51, 52, 64], we had sufficient sample size and sequencing depth to annotate the top six in the LBCs. We identified 73 participants (from 1,136) with CH (6%; Figure 3.1A). The gene-specific prevalence ranged between 1-36 cases with CH-variant allele frequencies ranging from 0.034-0.677 (Figure 3.1B). Mutations in *TET2* were exclusively frameshift and mutations detected in *JAK2* (all V617F), *SF3B1* and *TP53* were exclusively missense.

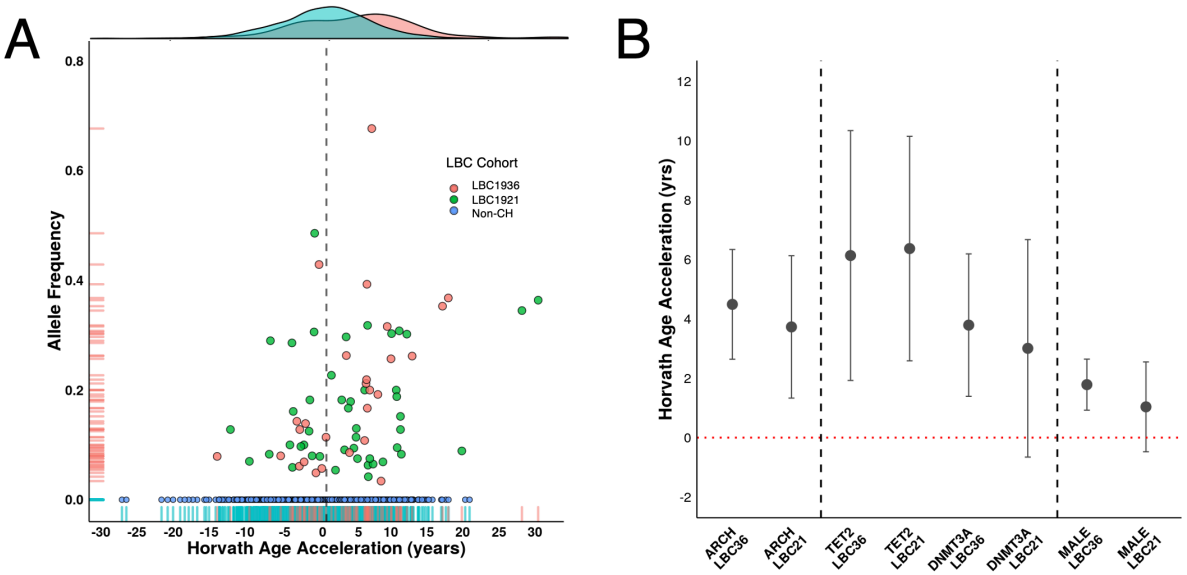


Figure 3.2: Effect of clonal haematopoiesis on epigenetic age estimates in the IEAA (Horvath) clock. ***A.*** *Scatter plot of Horvath age acceleration (IEAA; years) for individual LBC participants against the allele frequency of their CH variant in both LBC1921 (orange dots, net 3.7 years; p = 2.5 x 10^{-3}) and LBC1936 (green dots, net 4.5 years; p = 2.3 x 10^{-6}) cohorts. Density plot highlighting the shift in distribution of Horvath age acceleration between CH-positive (orange) and -negative participant (blue) groups. Non-CH carriers (blue dots).* ***B.*** *Plot showing net IEAA in CH (with 95% confidence intervals). The effect of sex (male versus female) on epigenetic ageing within the LBC is shown for comparison.*

CH status was associated with a significant increase in Horvath age acceleration: the increase was 4.5 (SE 0.9) years in LBC1936 and 3.7 (SE 1.2) years in LBC1921 (p = 2.3 x 10^{-6} and 2.5 x 10^{-3}, respectively; Figure 3.2A and Table 3.2). Compared with non-CH carriers, those with TET2 mutations had a 6.1 (SE 2.2) year and 6.4 (SE 1.9) year increase in Horvath age acceleration in LBC1936 and LBC1921 (p = 0.004 and p = 0.001), respectively. Those with DNMT3A mutations had 3.8 (SE 1.2) years increase in LBC1936, and 3.0 (SE 1.9) years in LBC1921 (p = 0.002 and p = 0.11), respectively (Figure 3.2B). These effect sizes are much larger than the sex-based differences in Horvath age acceleration, which were 1.8 (SE 0.4) years for men in LBC1936 (p = 5.1 x 10^{-5}), and 1.0 years (SE 0.8) in LBC1921 (p = 0.18) (Figure 3.2B and Table 3.2).

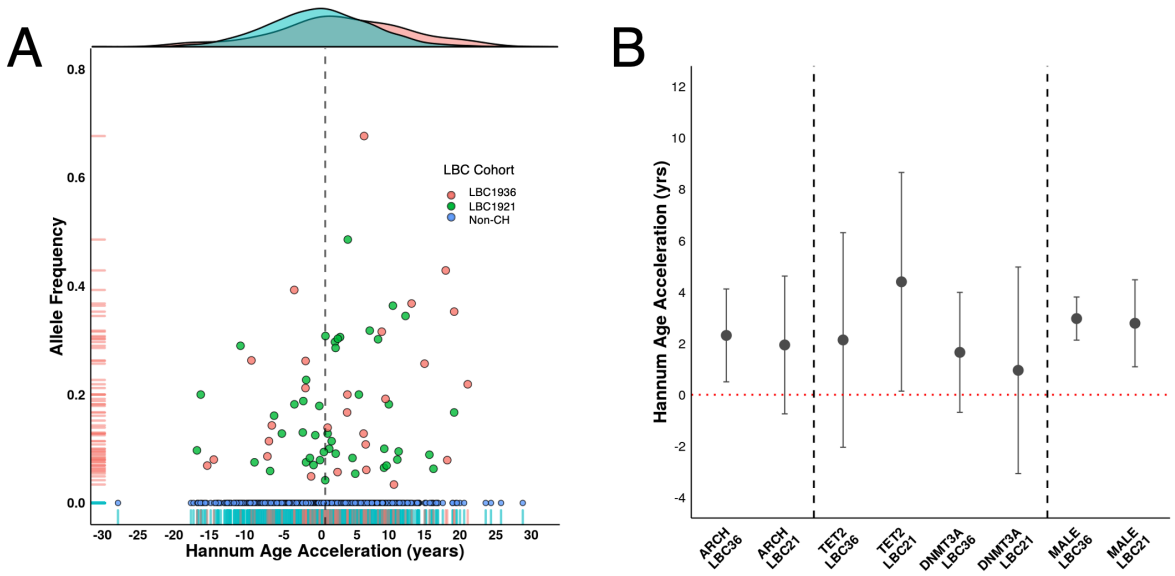


Figure 3.3: Effect of clonal haematopoiesis on epigenetic age estimates in the EEAA (Hannum) clock. ***A.*** *Scatter plot showing the Hannum age acceleration (EEAA; years) against the allele frequency of CH variants in both LBC1921 (orange dots, net 1.9 years; p = 0.16) and LBC1936 (green dots, net 2.3 years; p = 0.01) cohorts. Density plot highlighting shift in distribution of EEAA between CH-positive (orange) and -negative participant (turquoise) groups. Non-CH carriers (blue dots).* ***B.*** *Plot showing the net EEAA in CH (with 95% confidence intervals). The effect of sex (male versus female) on epigenetic ageing within the LBC is shown for comparison.*

We also considered age acceleration estimates from four additional epigenetic clocks: Extrinsic (Hannum) Epigenetic Age (EEAA) [249], PhenoAge [229], GrimAge [247] and Zhang Age [250] (Figure 3.3A,B and Figure 3.4A–F). Briefly, CH status was linked to increased EEAA, PhenoAge, GrimAge and ZhangAge, acceleration in LBC1921 (effect sizes: 1.9 years, 3.7 years, 2.8 years and 0.8 years with p = 0.16, 0.014, 9.6 x 10^{-4}, and 3.5 x 10^{-3}, respectively). In LBC1936 there was a modest association between CH and increased EEAA and ZhangAge (2.3 years and 0.5 years, p = 0.012 and 4.4 x 10^{-3}) but no association with PhenoAge or GrimAge acceleration (p = 0.32 and 0.99, respectively).

	LBC1936			LBC1921		
CH	**Beta**	**SE**	**P-Value**	**Beta**	**SE**	**P-Value**
IEAA	4.49	0.94	2.29E-06	3.73	1.22	2.54E-03
EEAA	2.31	0.92	1.24E-02	1.94	1.37	1.57E-01
Zhang	0.45	0.16	4.40E-03	0.76	0.26	3.46E-03
PhenoAge	1.06	1.06	3.15E-01	3.65	1.48	1.43E-02
GrimAge	0.01	0.71	9.89E-01	2.81	0.84	9.55E-04
TET2						
IEAA	6.13	2.15	4.38E-03	6.37	1.93	1.11E-03
EEAA	2.14	2.13	3.17E-01	4.40	2.17	4.39E-02
Zhang	0.66	0.37	7.75E-02	1.20	0.40	2.72E-03
PhenoAge	2.75	2.41	2.56E-01	3.20	2.31	1.68E-01
GrimAge	-0.77	1.65	6.42E-01	2.98	1.26	1.88E-02
DNMT3A						
IEAA	3.79	1.22	2.02E-03	3.01	1.87	1.09E-01
EEAA	1.65	1.19	1.66E-01	0.95	2.05	6.43E-01
Zhang	0.40	0.21	6.00E-02	0.35	0.37	3.49E-01
PhenoAge	0.55	1.37	6.88E-01	3.85	2.23	8.49E-02
GrimAge	-0.23	0.93	8.02E-01	1.56	1.22	2.02E-01
Sex (Male)						
IEAA	1.78	0.44	5.06E-05	1.04	0.77	1.80E-01
EEAA	2.97	0.43	8.53E-12	2.78	0.86	1.43E-03
Zhang	0.19	0.07	9.16E-03	0.17	0.16	2.90E-01
PhenoAge	1.85	0.49	1.73E-04	-0.56	0.93	5.48E-01
GrimAge	3.68	0.33	2.00E-16	2.85	0.53	1.76E-07

Table 3.2: A summary of associations with clonal haematopoiesis and epigenetic age estimates. Alongside this, specific CH genes are represented where sufficient mutational prevalence is achieved.

	LBC1936				LBC1921			
CH	**OR**	**LCI (95%)**	**UCI (95%)**	**P**	**OR**	**LCI (95%)**	**UCI (95%)**	**P**
Age	1.13	0.78	1.64	0.54	1.07	0.95	1.22	0.28
Sex (M)	1.06	0.57	2.02	0.85	0.60	0.23	1.45	0.27
NK	0.57	0.37	0.84	0.01	1.17	0.74	1.80	0.48
Mono	1.26	0.91	1.73	0.16	1.12	0.73	1.72	0.59
B-cell	1.10	0.86	1.31	0.34	1.37	1.01	1.94	0.05
CD4T	0.85	0.61	1.15	0.30	0.65	0.40	1.01	0.07
CD8T	0.77	0.46	1.17	0.28	0.94	0.58	1.42	0.79
TET2								
Age	1.54	0.62	4.08	0.36	1.13	0.93	1.44	0.25
Sex (M)	1.33	0.31	6.87	0.71	0.88	0.18	3.99	0.87
NK	0.52	0.17	1.26	0.20	1.27	0.65	2.34	0.46
Mono	1.78	0.85	3.51	0.11	1.79	0.92	3.65	0.09
B-cell	1.04	0.33	1.45	0.89	1.43	0.94	2.10	0.06
CD4T	0.95	0.43	1.97	0.90	0.68	0.30	1.38	0.32
CD8T	0.39	0.02	1.52	0.35	1.36	0.72	2.34	0.29
DNMT3A								
Age	1.22	0.74	2.03	0.44	0.99	0.83	1.20	0.92
Sex (M)	0.59	0.25	1.35	0.22	0.15	0.01	0.84	0.08
NK	0.56	0.32	0.93	0.04	0.57	0.24	1.22	0.18
Mono	1.61	1.07	2.38	0.02	1.03	0.49	2.19	0.93
B-cell	0.83	0.34	1.23	0.60	1.29	0.85	1.89	0.17
CD4T	0.92	0.59	1.45	0.72	0.77	0.39	1.42	0.43
CD8T	1.03	0.58	1.60	0.92	1.05	0.40	2.06	0.90

Table 3.3: Associations of blood cell count proportions with CH status, CH specific genes and sex (male versus female).

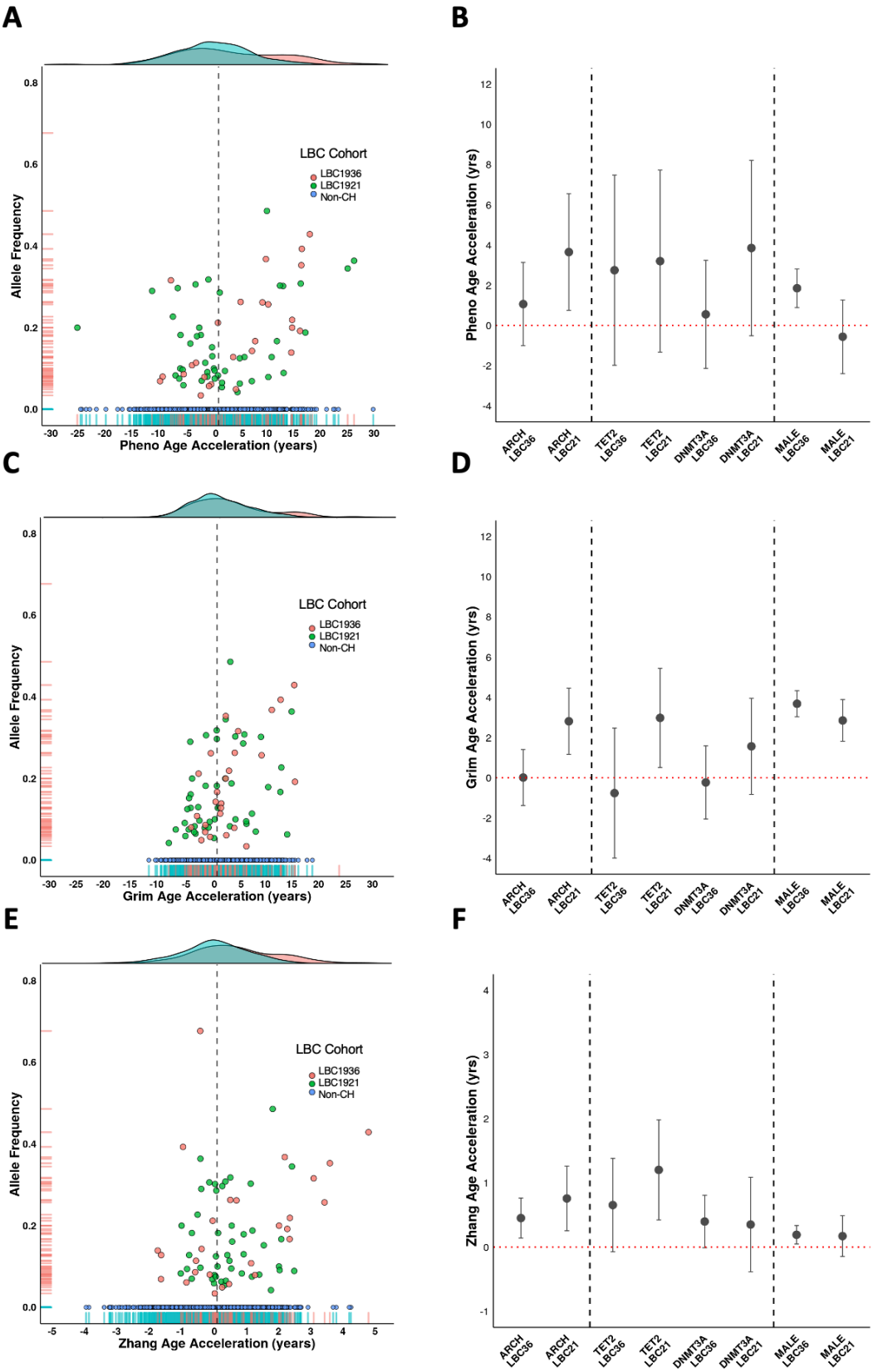


Figure 3.4: Effect of clonal haematopoiesis on epigenetic age estimates in the PhenoAge, GrimAge and ZhangAge clocks. *Legend as Figure 3.2 and Figure 3.3 for the PhenoAge (A & B), GrimAge (C & D) and ZhangAge (E & F) clocks, respectively.*

There was no consistent association between CH status and white cell count proportions across the two cohorts: a lower proportion of NK cells was linked with CH carrier status in LBC1936 (odds ratio per SD of cell counts, 0.57 95% CI [0.37, 0.84]), while a higher B cell proportion was associated with CH status in LBC1921 (OR 1.37 [1.01, 1.94]) (Table 3.3) [4].

3.3 Conclusion and Discussion

In these results, we observed significant associations between CH and epigenetic age acceleration in the independent Lothian Birth Cohorts of 1921 and 1936 in a variety of epigenetic clocks, though not all. Firstly, we observed the strongest associations with CH in the canonical clocks IEAA, EEAA and ZhangAge – with the intrinsic clock, IEAA, being particularly effective. Earlier studies into the potential mechanisms of intrinsic age-acceleration highlighted its correlation with the number of population doublings in common human cell-lines [284, 285]. This might imply that the strong response from the intrinsic clocks to CH may be driven by heightened levels of self-renewal or proliferation in HSPC clones in cells with corresponding driver mutations, or conversely, exhaustion of the WT HSPCs leading to the positive section of mutant clones [284, 285]. It is possible that we might be able to delineate these effects in certain model systems by longitudinally assessing clock rates in tandem with accurate clonal VAF measurements. In addition, the ZhangAge intrinsic clock also showed significant associations with CH in both the LBC1921 and the LBC1936.

The extrinsic ageing clocks showed more modest associations with CHIP (2.3 years in the LBC1936 and 1.9 years in LBC1921, $p = 0.012$ and $p = 0.15$, respectively). Individuals regularly exhibit reductions in lymphoid cell counts in old age that results in a reduction in immune function [32]. It's possible that a more general age-dependent shift in blood cell count proportions, particularly in the older LBC1921 cohort, masked the effects of CHIP in our analysis. While we observed minimal associations with the composite clocks (PhenoAge and GrimAge) in the younger LBC1936 cohort, the older LBC1921 cohort exhibited significant associations in both. This might be a facet of the

long timeframes required for CH to become pathologically relevant, or driven by selection biases in the composition of the older cohort.

We also observed significant associations with DNMT3A and TET2 mutations in the intrinsic ageing clocks. Due to the lack of mutational coverage, no other genes associations were measured. In future, larger cohorts will be required to achieve a more complete appraisal of the links of CH to epigenetic ageing across a more diverse range of genetic drivers. In addition, one might look for improvements in the measurement of VAF levels for these drivers via error-corrected or even exome sequencing to better understand its links to clone size.

After this work was published in 2019, the links between CH and epigenetic ageing were reciprocated in three further studies. Nachun et al. showed that CH, DNMT3A and TET2 are significantly associated with accelerated epigenetic age at levels consistent with ours. Thanks to their larger cohort, they were able to probe deeper into the genetic background of CH and highlighted that individuals with multiple genetic drivers display the greatest age-accelerations [285]. Thereafter, Feldkamp et al. showed that the VAF of somatic CH mutations have a significant impact on epigenetic age estimates, highlighting the importance of clone size in CH - orthologous to the work of Abelson who showed that clone size has a significant impact on cancer progression and disease risk [3, 286]. Finally, Soerensen et al. used a small cohort of Danish twins and found significant links to clone size and eAge across the cohort and between twin pairs [287].

To conclude, these results show an important relationship between CH and epigenetic ageing estimates.

Chapter 4: Longitudinal Dynamics of Clonal Haematopoiesis Identifies Gene-Specific Fitness Effects

Clonal haematopoiesis of indeterminate potential (CHIP) increases rapidly in prevalence beyond age 60 and has been associated with increased risk for malignancy, heart disease and ischemic stroke. CHIP is driven by somatic mutations in hematopoietic stem and progenitor cells (HSPCs). Since mutations in HSPCs often drive leukaemia, we hypothesised that HSPC fitness substantially contributes to transformation from CHIP to leukaemia. HSPC fitness is defined as the proliferative advantage over cells carrying no or only neutral mutations. If mutations in different genes lead to distinct fitness advantages, this could enable patient stratification. We quantified the fitness effects of mutations over 12 years in older age using longitudinal sequencing and developed a filtering method that considers individual mutational context alongside mutation co-occurrence to quantify the growth potential of variants within individuals. We find that gene-specific fitness differences can outweigh inter-individual variation and therefore could form the basis for personalised clinical management.

4.1 Introduction

Age is the single largest factor underlying the onset of many cancers [288]. 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 [289]. However, the widespread presence of cancer-associated mutations in healthy tissues highlights the complexity of early detection and diagnosis of cancer [1, 2, 191, 290, 291].

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 [1, 2, 290]. There is a clear benefit of detecting CHIP early for close clinical monitoring and early detection as the association between clone size and malignancy progression is well-established [1, 53, 292].

The particular mechanisms by which common mutations of CHIP, e.g., *DNMT3A* and *TET2* contribute to the progression of leukaemia are still not understood, which hinders early diagnosis of CHIP on a gene or variant-basis [53, 124, 293, 294]. In clinical practice, CHIP is diagnosed by the presence of somatic mutations at variant allele frequencies (VAF) of at least 2% in cancer-associated genes in more than 4% of all blood cells [1, 200]. 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 [202, 295]. As mutations in stem cells often drive leukaemia [1], we hypothesise that stem cell fitness contributes substantially to transformation from CHIP to leukaemia.

Stratification of individuals to inform close clinical monitoring for early detection or prevention of leukaemia in the future 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 outweigh 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 [3, 202]. 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) [253]. 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 using Bayesian inference. Firstly, a likelihood-based filter for time-series data (LiFT) allowed us to segregate between sequencing artefacts or naturally drifting populations of cells and fast-growing clones. Secondly, we infer the growth potential or fitness effects simultaneously for all growing mutations within each individual and also allow for clones with multiple mutations if these are favoured by Bayesian model comparison. We detected gene-specific fitness effects within our cohorts, highlighting the potential for personalised clinical management.

4.2 Results

4.2.1 Longitudinal Profiling of CH 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) [253]. We previously identified 73 participants with CHIP at Wave 1 through whole-genome sequencing (WGS) [4]. Here, we used a targeted error-corrected sequencing approach using a 75 gene panel (ArcherDX/Invitae; Table 2.1) to assess longitudinal changes in variant allele frequencies (VAF) and clonal evolution over 21 years across both LBC cohorts (6 years in LBC21 and 12 years in LBC36; Table 2.2).

Error-corrected sequencing allowed accurate quantification, providing more sensitive clonal outgrowth estimates compared to our previous WGS data. We sequenced 248 LBC samples (85 individuals across 2-5 time-points) and achieved a sequencing depth of 2238x mean coverage (2153x median) over all targeted sites with an average of 1.6 unique somatic variants (pan-cohort VAF 0.03-87%, median VAF 4.4%) detected per participant. We examined all participant-matched events across the time-course: sequence quality control metrics revealed that only 7 of 275 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 258 (Figure 2.1A).

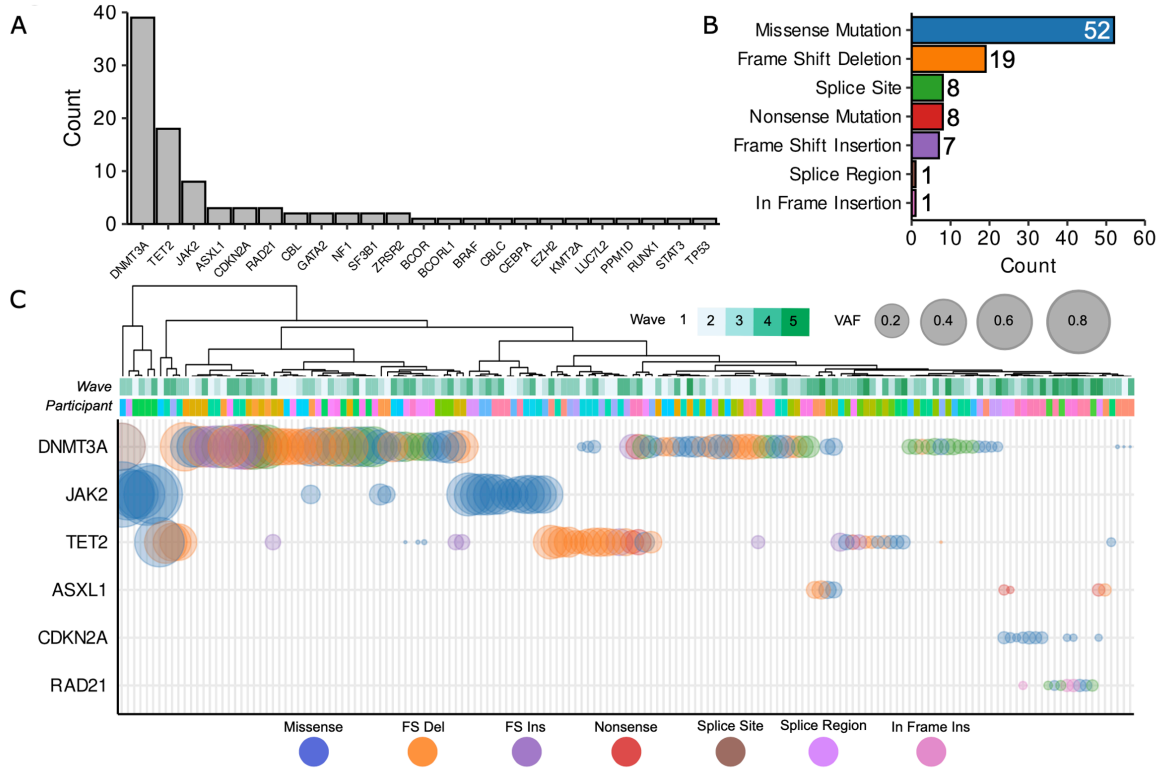


Figure 4.1: Unique clonal haematopoiesis variants at 2% VAF in the LBCs. *A.* *Counts of unique events that exceeded 2% VAF across the range of the longitudinal cohorts in our panel of 75 hematopoietic genes.* ***B.*** *Counts of the functional consequences of the unique events listed in Figure 4.1A. Missense mutations, frameshift insertions and deletions and nonsense mutations are indicated. Exact counts, n, are for each category.* ***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 were clustered across all timepoints, with the point size scaled by VAF and coloured by the functional consequence of the variant (as per Figure 4.1B and legend). Key: del, deletion; FS, frameshift; ins, insertion.*

For our initial analysis, we retained variants with at least one time-point at 2% VAF (Appendix 1). *DNMT3A* was the most commonly mutated CHIP gene (n=39 events in 33 participants), followed by *TET2* (n=18 events in 15 participants), *JAK2* (n=8 events in 8 participants) and *ASXL1* (n= 3 events in 3 participants) (Figure 4.1A-C; Figure 4.2). Our mutation spectrum is consistent with previous studies in finding *DNMT3A* and *TET2* as the most frequently mutated genes [1, 2]. We detected some variants more frequently at certain hot spots within a gene such as R882H in *DNMT3A*, with previously unreported variants being present as well (Figure 4.3A-F, Appendix 1) [1]. We most frequently detect missense mutations with several other key protein altering

event types ranking highly, including frameshift insertions/deletions and nonsense mutations (Figure 4.1A-C).

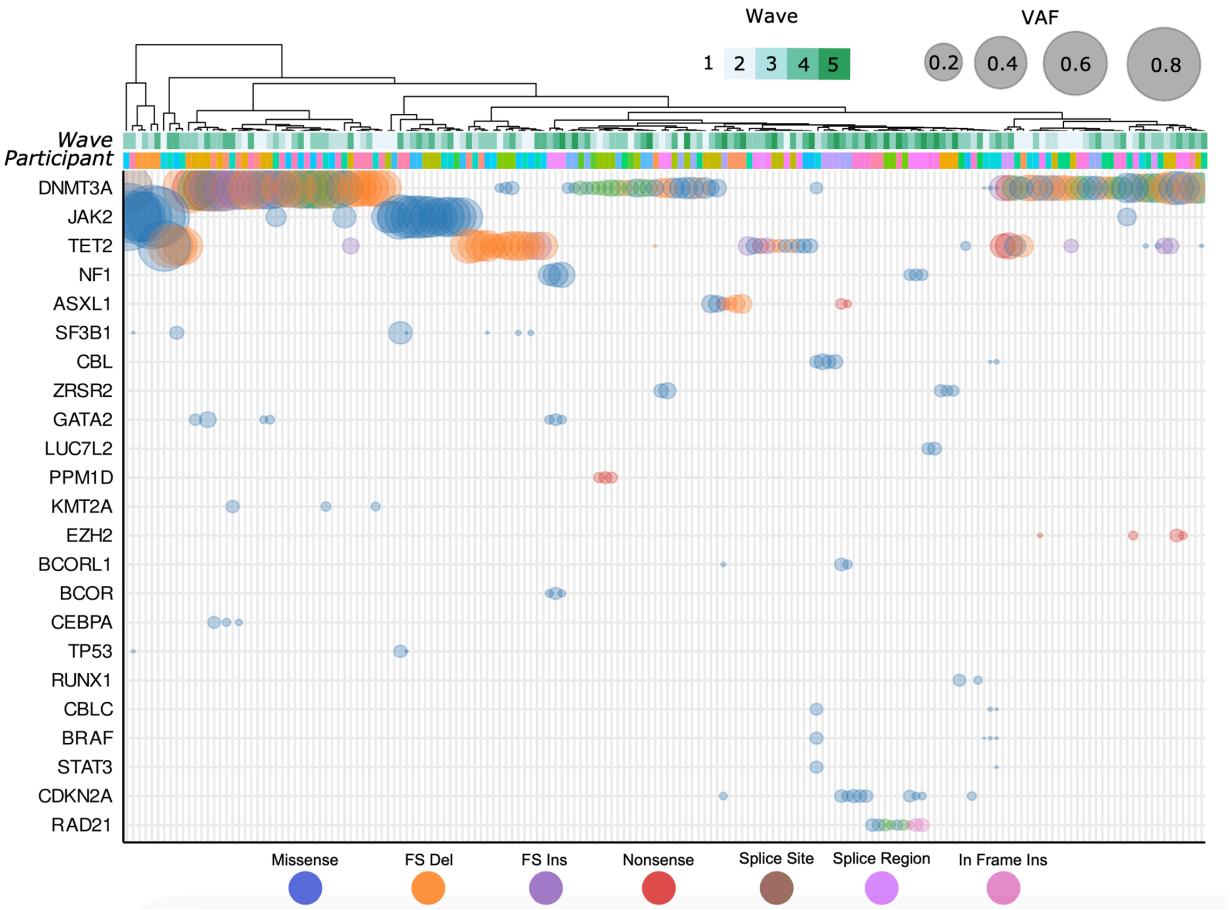


Figure 4.2: Heatmap of all captured variants across all timepoints. *Schematic of all affected genes in the cohort with the largest clone size of an event in any given gene shown above 2% VAF. All affected participants have been clustered across all time-points, with the point size scaled by VAF and coloured by the functional consequence of the variant (as per legend). Key: del, deletion; FS, frameshift; ins, insertion.*

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* (Figure 4.1A-C). 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 (Figure 4.3E). Those individuals were excluded from further analysis. In our data, we identified a lower frequency of

mutations in splicing genes, such as *SF3B1*, despite the older age of the cohorts (Figure 4.1A and Figure 4.2). This is in contrast to previously published cohort data, where splicing mutations became more prominent with increased age [52]. The majority of mutations were missense, frameshift and nonsense mutations (Figure 4.1B).

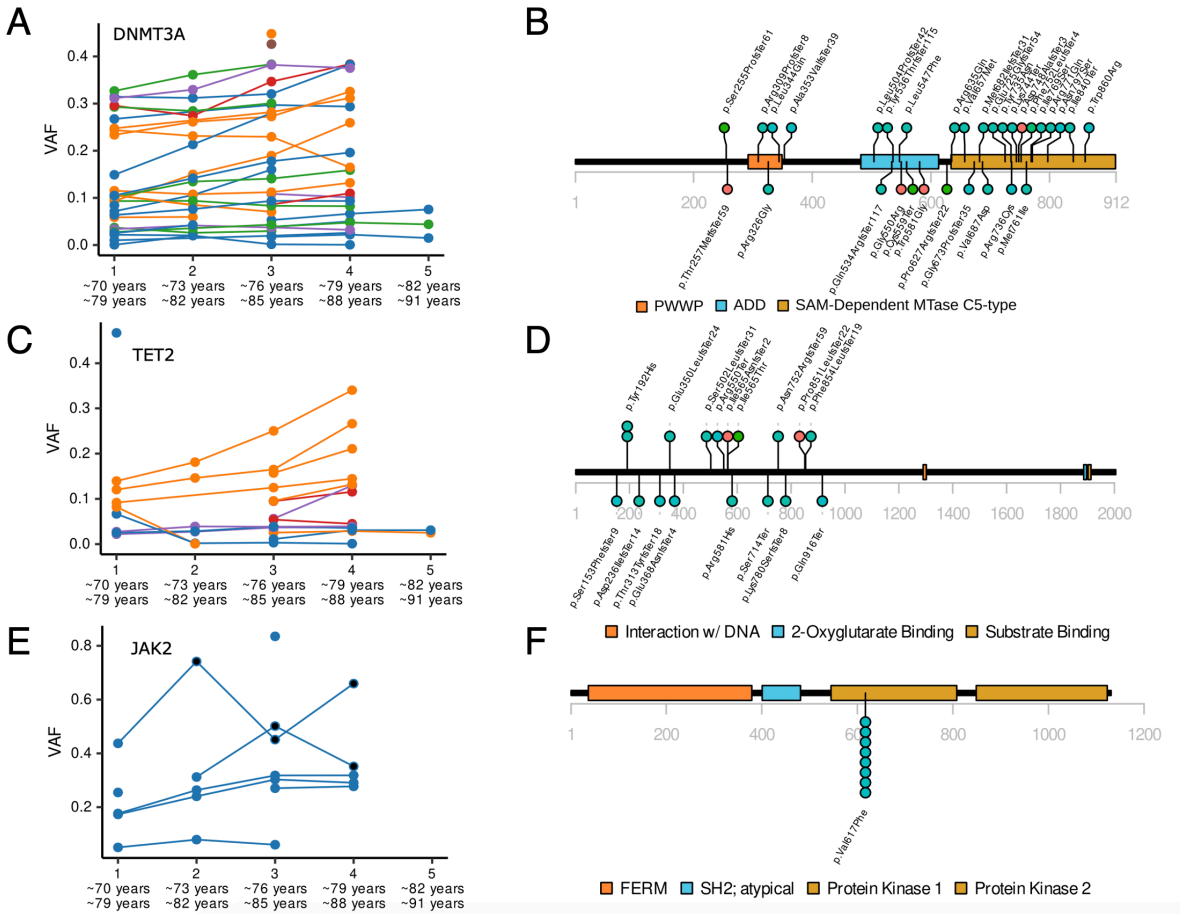


Figure 4.3: VAF trajectories across the time-series alongside the locations of protein affecting mutations in DNMT3A, TET2 and JAK2. ***A.*** *Clone size trajectories of all DNMT3A mutations across the time series in both LBC1921 and LBC1936 coloured by the functional consequence of the variant (as per Figure 4.1A and 4.1C).* ***B.*** *Locations of somatic mutations discovered in DNMT3A. Protein-affecting events are marked and labelled across the structure of the gene (missense in red, truncating in purple, stacked for multiple events) with the structure of the gene labelled along the amino acid length of its protein.* ***C.*** *Clone size trajectories of all TET2 mutations across the time series in both LBC1921 and LBC1936 coloured by the functional consequence of the variant* ***D.*** *The locations of somatic mutations in TET2. Protein-affecting events are marked and labelled across the structure of*

*the protein **E.** Clone size trajectories of all JAK2 mutations across the time series in both LBC1921 and LBC1936 coloured by the functional consequence of the variant. Points marked in black denote timepoints after which the affected participant received treatment for leukaemia. **F.** The locations of protein-affecting somatic events are marked and labelled across the structure of the JAK2 protein. All eight JAK2 mutations are p.Val617Phe (JAK2 V617F) missense variants.*

Overall, our sequencing approach allowed for high resolution, longitudinal mapping of CHIP variants over 6 and 12-year time spans in the LBC21 and LBC36, respectively, and 21-year time-span across both cohorts from the same geographical region and born 9 years apart.

4.2.2 Cataloguing the Fitness Effects for CH Variants at >2% VAF

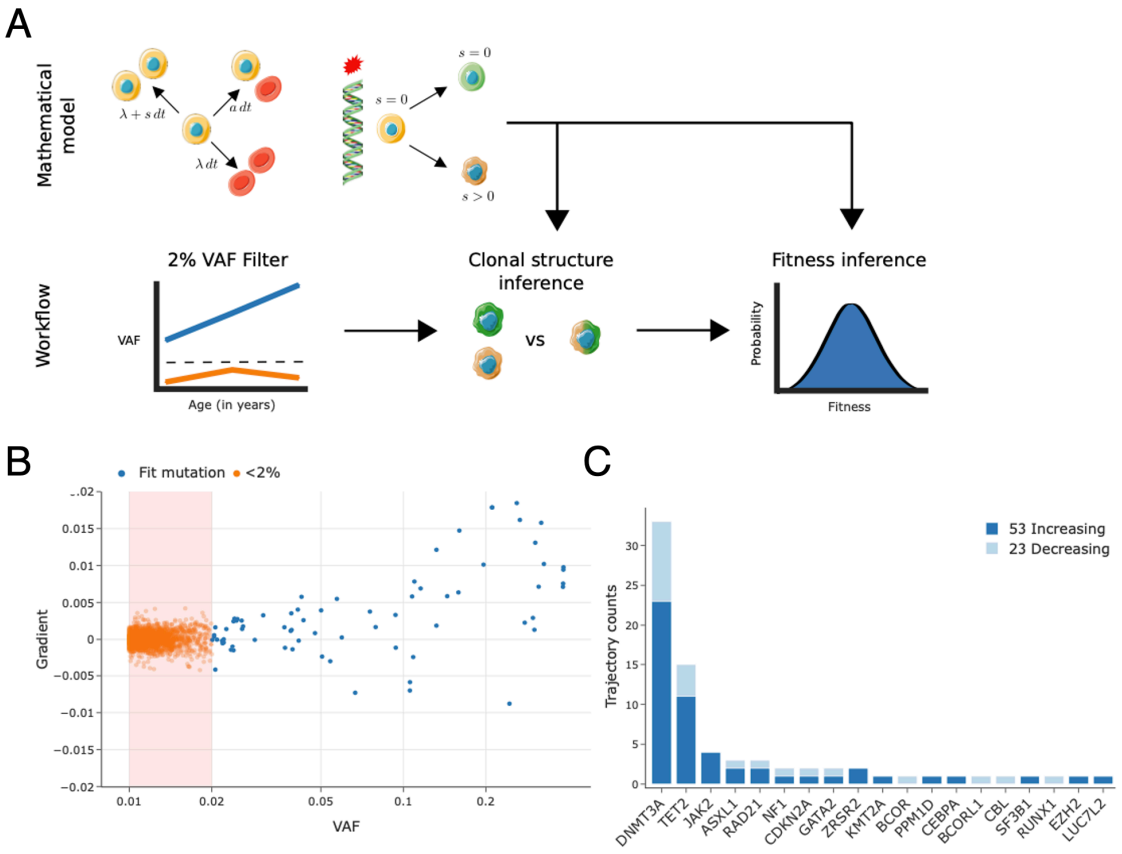


Figure 4.4: Model to capture the gradients and growth potential of variants at 2% VAF threshold in longitudinal data. ***A.*** *Schematic of the mathematical model (top)*

and workflow (bottom) used to infer the fitness of mutations reaching VAF>2% during the observed time span. Clonal structure and fitness inference are based on a mathematical model of clonal dynamics (Methods). HSPCs (top, yellow cells) naturally acquire mutations over time that can be neutral (s=0, green cell) or increase self-renewal bias (s>0, brown cell), leading to the formation of genetic clones. ***B.*** *VAF measurement v(t_0) at initial timepoint t_0 versus gradient in VAF, $(v(t_{end})-v(t_0))/(t_{end}-t0)$, between initial and last timepoints t_0 and t_{end} of all variants detected in the LBCs with at least two timepoints. Each data point corresponds to a trajectory in the LBCs and has been coloured according to its CHIP status based on the 2% VAF threshold (red box). Blue and orange, respectively, denote whether trajectories achieved a VAF>2% during the observed time span or not. Note: VAF is displayed on a logarithmic scale, as most mutations are concentrated at low VAF.* ***C.*** *Number of trajectories passing the currently used 2% VAF threshold, broken down into whether VAF is increasing or decreasing from the first to last timepoint. Artwork includes images by Servier Medical Art licensed under CC BY 3.0.*

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 [202], which would make it difficult to clinically stratify individuals with CHIP.

To determine the fitness effects of the variants identified in our cohorts (Figure 4.1A; Figure 4.2), we initially selected all CHIP variants in our data using the commonly used criterion of defining any variants with VAF>2% as CHIP [53, 200], and retaining only those variants with at least 2 time-points (Figure 4.4B). This approach identified 76 CHIP mutations overall (Figure 4.4C). To estimate the fitness effect each variant confers, we use Bayesian inference and birth-death models of clonal dynamics (Figure 4.4A) including all trajectories with at least 2 time-points (Appendix 2). The resulting fitness values show an overall dependence of fitness on the gene level (Figure 4.5), with a wide distribution of fitness for some genes, such as *TET2* and *DNMT3A*, but not others such as *JAK2* (which are all the same variant).

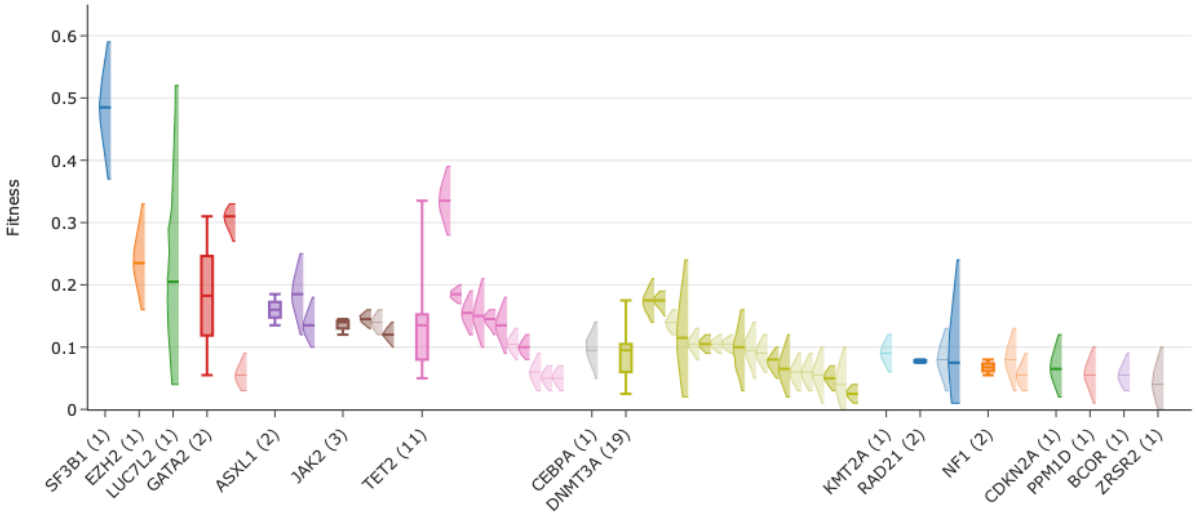


Figure 4.5: The fitness effects of variants at the 2% VAF threshold. *Fitness effects of mutations grouped by gene and ranked by median fitness. The posterior probability distribution of the fitness as inferred from our model of clonal dynamics is displayed for each mutation (only the 90% interval is shown). The sample size, n, of observed variants in each gene is denoted in brackets. When more than one mutation is observed in a gene, we further display a box plot showing the median and exclusive interquartile range of the MAP fitness estimates associated with the gene.*

4.2.3 Longitudinal Trajectories Accurately Stratify 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>2% threshold not only misses fast growing and potentially harmful variants (Figure 4.4B) but can include variants whose frequencies are shrinking (Figure 4.4B, 4.4C) and thus either do not confer a fitness advantage or are being outcompeted by other clones. Overall, 70% of CHIP mutations detected by thresholding at 2% VAF were growing during the observed time span (Figure 4.4B, 4.4C). Longitudinal data thus reveal limitations in defining CHIP mutations based on a widely used VAF threshold.

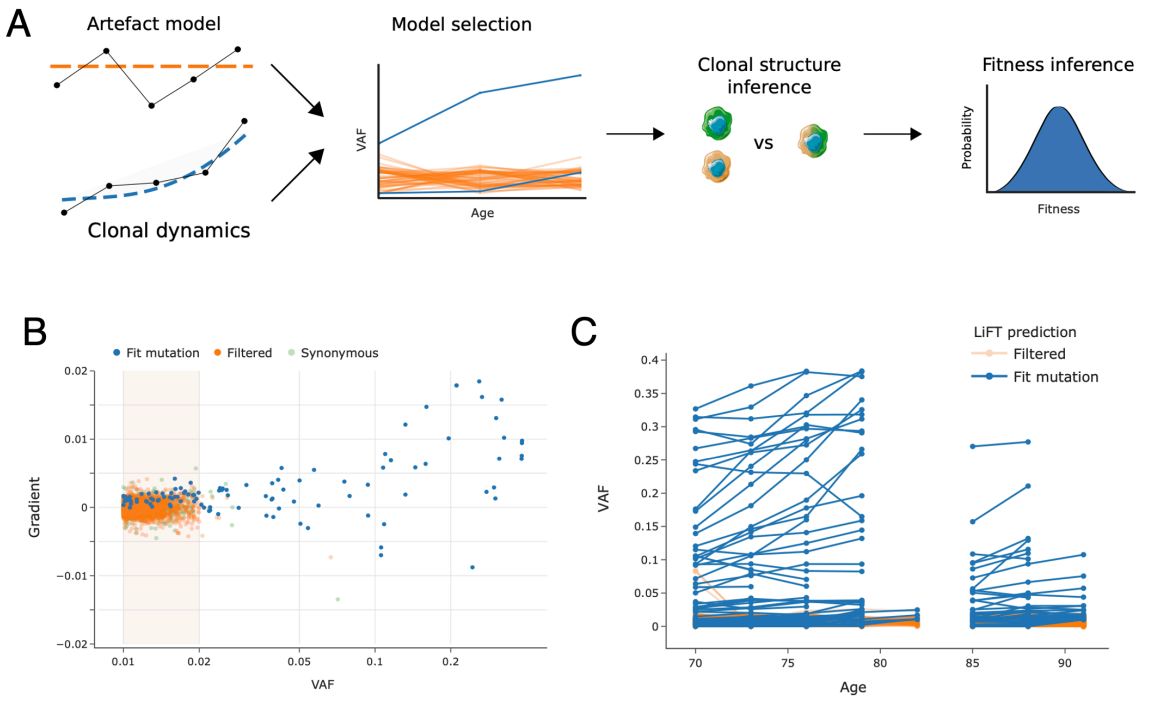


Figure 4.6: LiFT allows for the classification of fit variants <2% VAF. ***A.*** *Schematic of LiFT algorithm. LiFT compares a model of clonal dynamics with an artifact model and performs Bayesian model selection. The subsequent steps to infer clonal structure and fitness distributions are as in Figure 4.4A and 4.4B.* ***B.*** *Gradient in VAF versus VAF for variants detected in the LBCs with at least two timepoints and at least one VAF>1% per trajectory, with filtered (orange), fit (blue) and synonymous (light green dots) mutations, classified by LiFT on a logarithmic scale.* ***C.*** *Longitudinal trajectories of fit (blue) and filtered (orange) mutations linked to age in years.*

To overcome the limitations of a threshold-based selection of fit variants, we sought to filter variants based on longitudinal information, by comparing a stochastic model of clonal dynamics with a model of sequencing artefacts (Figure 4.6A). This novel approach, which we named Likelihood-based filter for time-series data (LiFT), allows classification of fit variants even for VAF<2%. LiFT classification of fit variants broadly agreed with noise profile statistics from the Archer DX pipeline (Figure 2.2A, 2.2B), but identified additional variants by leveraging the longitudinal nature of the data.

LiFT classification resulted in 114 variant trajectories (Figure 4.6B-D), 86% 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 2% VAF, with only 70% of variants identified to be

growing and thus likely to confer a fitness advantage. Of 114 variants we detected, 50 would not have been detected using the previous VAF-threshold filter. We therefore recomputed fitness estimates for this new set of fit trajectories (Figure 4.7B, 4.7C). Growing variants that were missed by the traditional filtering method include highly fit variants such as *U2AF1* Q157R (fitness 33.5%) and *DNMT3A* R882H (fitness 16%) (Figure 4.6C, 4.7D, Appendix 3). VAF-thresholding did not identify any *TP53* variants. However, LiFT identified four *TP53* mutations, all of which were growing over the observed time-course (Figure 4.6C, 4.7D, Appendix 3). In addition, all of those were either termination/frameshift mutations or were previously reported as cancer associated in COSMIC [259] and classified as likely damaging (Appendix 4). Moreover, all *TP53* variants led to high fitness effects, thus our filtering method allows us to identify potentially harmful variants at very low VAFs. Overall, the variants detected by LiFT were of higher fitness than those detected by VAF-thresholding (Figure 4.7C; Kruskal-Wallis H=14, $p=1 \times 10^{-4}$), with an even larger effect size when comparing variants that are exclusive to each filtering algorithm (Figure 4.7C; Kruskal-Wallis H=18, $p=1 \times 10^{-5}$).

We further stratified variants using seven computational predictors recently identified as being most useful for identifying pathogenic mutations [271–277] (Figure 4.7D and Appendix 4), categorising the most prevalent CHIP variants into likely damaging (21 variants), possibly damaging (20 variants) and likely benign (11 variants), as well as frameshifts and terminations (37 variants, which are also most likely damaging to protein structure and thus protein function). Our novel LiFT algorithm therefore produces a low false discovery rate of pathogenic variants, with 88% of the detected fit variants being predicted to be possibly damaging, frameshift or termination.

Taken together, applying a probabilistic model of clonal dynamics to longitudinal sequencing data results in a novel method, the LiFT algorithm, that improves on the threshold-based definition of CHIP mutations (Figure 4.6A). The LiFT algorithm replaces an arbitrary cut-off on VAF by a choice of false discovery rate (through a Bayes Factor threshold) and as a result selects fewer trajectories with shrinking VAF (Figure 4.4B, 4.4C and Figures 4.6B, 4.7A).

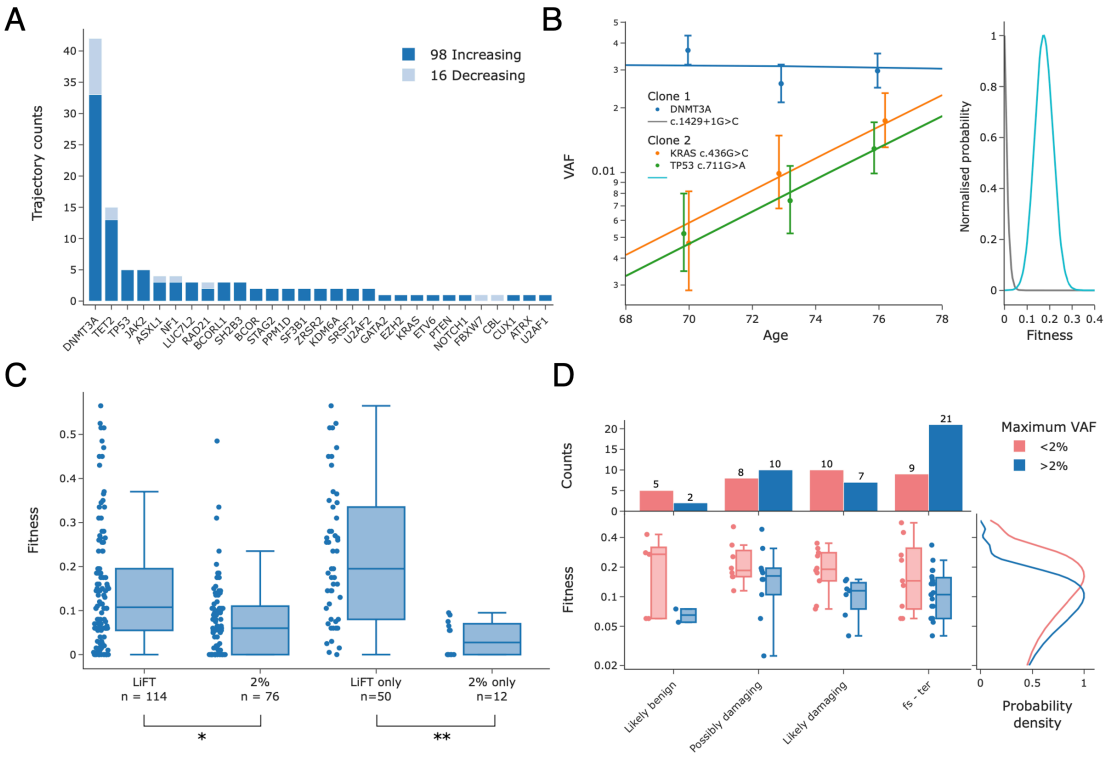


Figure 4.7: LiFT allows for the classification and inference of clonal structure of fit variants <2% VAF. ***A.*** *Number of trajectories classified as fit by LiFT, broken down into increasing or decreasing VAF from first to last timepoint.* ***B.*** *Left, deterministic fit of all mutations selected by LiFT in an individual of the LBC cohorts using the inferred optimal clonal structure. 90% CIs associated with binomial sampling noise are shown for each data point. VAF is displayed on a logarithmic scale. Right, posterior distribution of fitness associated to each clonal structure.* ***C.*** *Fitness effects of variants broken down by filtering method. The sample size, n, and statistical analyses comparing the distribution of fitness, computed using the non-parametric Kruskal–Wallis test, are highlighted (*H = 14, P = 1 × 10^{-4}; **H = 18, P = 1 × 10^{-5}).* ***D.*** *Fitness of variants selected as fit by LiFT broken down by their maximum VAF, >2% and <2%, and damage prediction. The top row displays a bar plot of variant counts for each category. The bottom row displays box plots showing the median and interquartile range of the distribution of MAP fitnesses by damaging prediction displayed on a logarithmic scale to emphasize relative differences in fitness between variants. Consequently, of a total of 89 variants with a damage prediction, 17 variants whose damage prediction was difficult to discern with fitness below 2% are not shown but are reported in Appendix 3. A marginal plot shows the Gaussian kernel density estimation of the MAP fitness values. Key: fs, frameshifts; ter, terminations.*

4.2.4 Clinical Relevance of LiFT

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 (Figure 4.8A, 4.8B). In particular, we found that mutations in *TP53, SF3B1 and SRSF2* conferred a higher fitness advantage over mutations in commonly mutated CHIP genes such as *JAK2* and *DNMT3A*. 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 and genes for transcription factors that are relevant in development (Figure 4.9A, 4.9B).

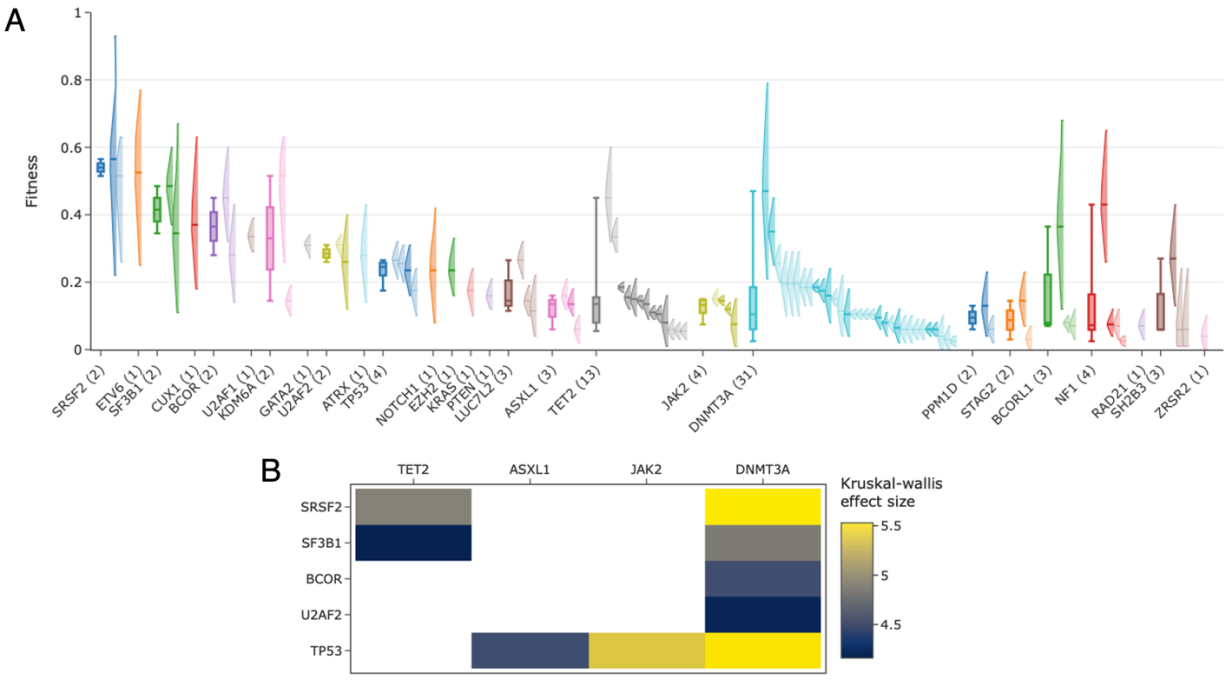


Figure 4.8: LiFT and gene specific fitness effects. ***A.*** *Fitness effects of mutations selected as fit by the LiFT algorithm, grouped by gene and ranked by median fitness. The posterior probability distribution of the fitness as inferred from our model of clonal dynamics is displayed for each mutation (only the 90% interval is shown). The sample size, n, of observed variants in each gene is denoted in brackets. When more than one mutation is observed in a gene, we further display a box plot showing the median and exclusive interquartile range of the MAP estimates of fitness associated with the gene.* ***B.*** *Analysis of variance of the distribution of fitness across genes. Heat map of all statistically significant (P<0.05) Kruskal–*

Wallis H statistics, labelled by effect size, computed for all combinations of pairs of genes. The effect size is only shown for statistically significant relations. Variants with a fitness below 2% were left out of this study, as our prediction classifies them as conferring no or a negligible fitness advantage.

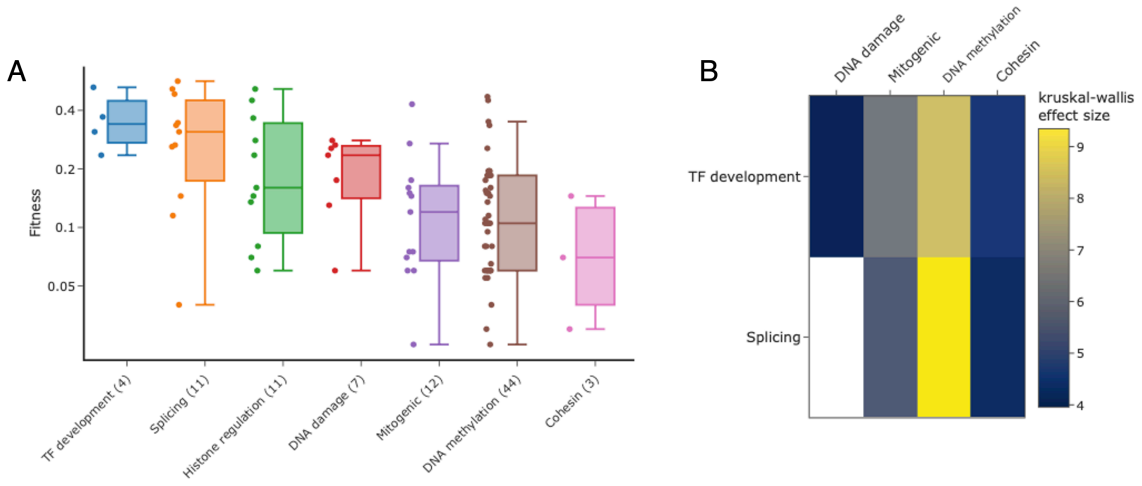


Figure 4.9: LiFT and gene-fitness summarised by ontological classes. ***A.*** *Distribution of fitness by gene category. Genes are grouped according to their biological function into DNA methylation (TET2, DNMT3A), Splicing (SF3B1, U2AF1, SRSF2, U2AF2, ZRSR2, LUC7L2, DDX41), mitogenic function (KRAS, NF1, JAK2, JAK3, SH2B3, PTEN, PTPN11, NRAS), cohesin (RAD21, STAG2), DNA damage (TP53, CDKN2A, PPM1D, ATRX) and Transcription factors (TF) important during development (GATA2, RUNX1, NOTCH1, CUX1, ETV6). The sample size, n, of each gene category is denoted in brackets. For each gene category we display a boxplot showing the maximum a posteriori (MAP) estimates of fitness for variants in the category, as well as the median and exclusive interquartile range.* ***B.*** *Analysis of variance of the maximum posterior fitness estimates across gene categories. Heatmap of all statistically significant ($p<0.05$) Kruskal-Wallis H statistics, labelled by effect size, computed for all combinations of pairs of genes. The effect size is only shown for statistically significant relations. Variants with a fitness below 2% were left out of this study as our prediction classifies them as conferring no or a negligible fitness advantage.*

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 a gene with 10% fitness at 1% VAF, its growth could be confidently measured after 7 months (Figure 4.10A), warranting a clinical follow-up over that time-frame to confirm

or revise the fitness estimate. Conversely, the time between observations places a lower bound on the fitness that can be measured for mutations of a given VAF (Figure 4.10B). These data can then inform on the time-frame for close clinical monitoring and early detection of disease.

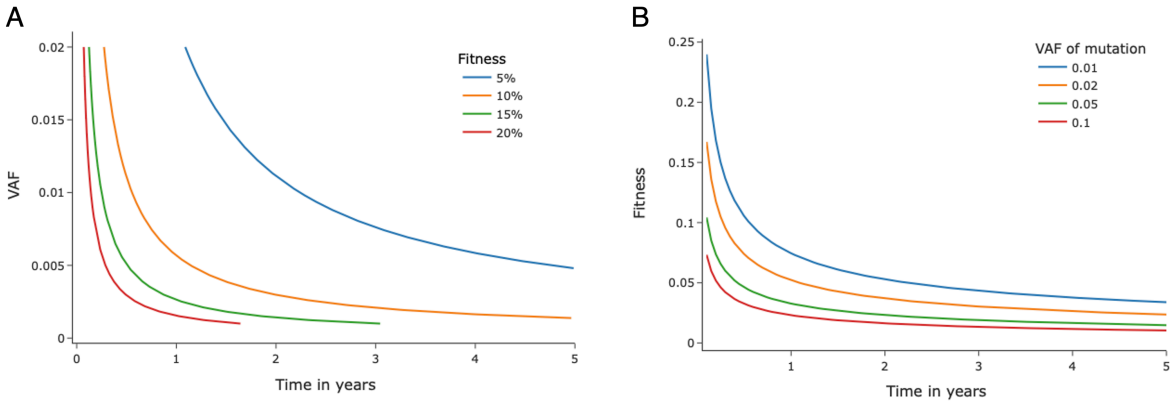


Figure 4.10: Clinical relevance of LiFT. ***A.*** *Minimum referral time in years based on 2 standard deviations below the expected growth of a clone given an initial VAF and fitness. Each line shows the initial size of mutation versus referral time for a given fitness.* ***B.*** *Minimum detectable fitness at referral observation based on 2 standard deviations below the expected growth of a clone given an initial VAF and fitness. Each line shows minimum detectable fitness versus referral time for an initial clone size.*

Abelson and colleagues compared CHIP carriers who never developed AML with CHIP where individuals subsequently developed AML and found that the number of mutations, the mutational burden and the size of the larger driver clone were associated with the risk of progression to AML [3]. In this study, we carried out a survival analysis to correlate the maximum observed VAF of mutations and survival. This correlation was stronger in the older cohort (LBC21), although not statistically significant (Hazard Ratio of 1.35, 95% CI [0.83, 2.19]; p=0.23) due to the small sample size (Figure 4.11, Table 4.1). In the younger cohort (LBC36), we found that survival better correlated with the speed of growth of a mutation, although this was again not statistically significant (Hazard Ratio of 1.35, 95% CI [0.76, 2.4]; p=0.3) (Figure 4.11, Table 4.1).

Importantly, only 2 time-points are necessary to apply LiFT, making this a widely applicable method for existing cohorts and future studies. We propose the use of LiFT over thresholding for clinical practice.

cohort	covariate	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
LBC21	growth speed	-0.199	0.819	0.245	-0.67	0.281	0.506	1.325	-0.81	0.417	1.2613
LBC21	max VAF	0.297	1.346	0.248	-0.19	0.785	0.826	2.193	1.19	0.231	2.1083
LBC36	growth speed	0.302	1.353	0.292	-0.27	0.876	0.762	2.402	1.03	0.300	1.7339
LBC36	max VAF	0.038	1.039	0.326	-0.60	0.678	0.548	1.971	0.11	0.905	0.1437

Table 4.1: Survival analysis on the effects of maximum VAF and clone growth speed. *Value p corresponds to the chi-squared test in the Cox hazard analysis.*

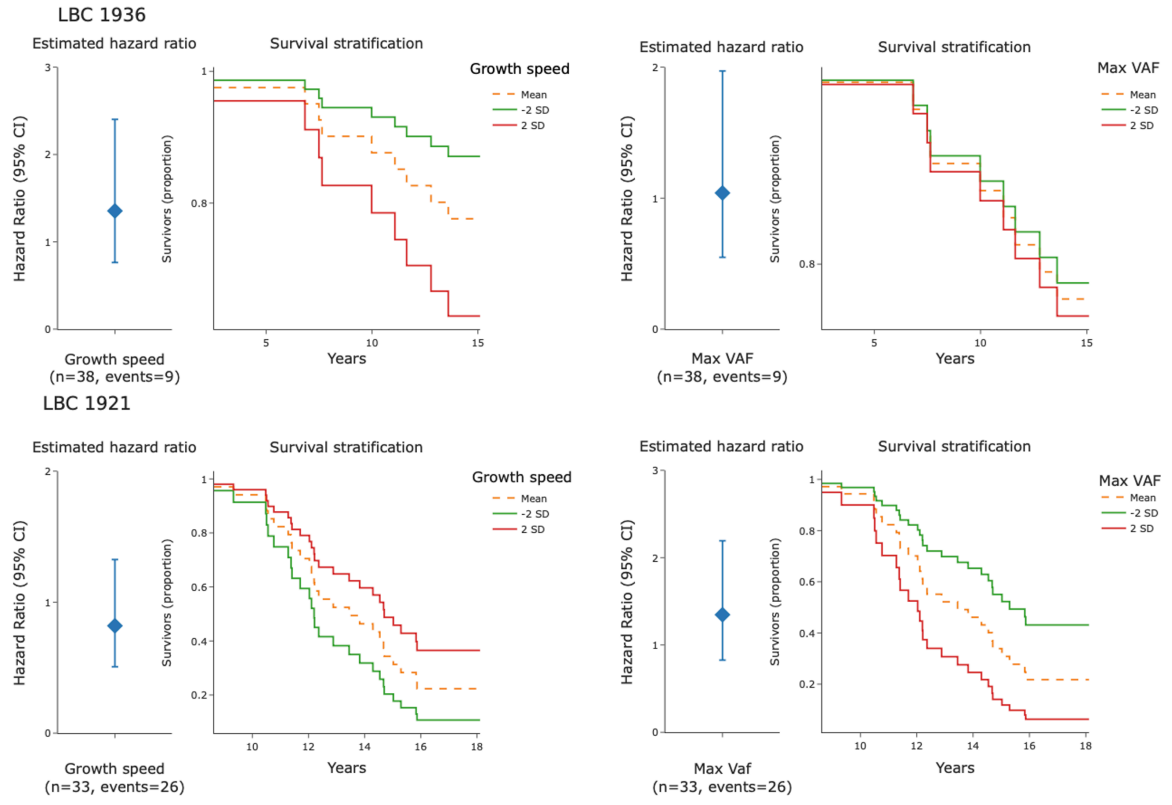


Figure 4.11: Estimations of gene fitness have the potential to provide a novel route to estimating clinical outcomes. *A.* *Survival analysis (Cox proportional hazards regression model) broken down by cohort and covariates. LBC1921 and LBC1936 are analysed separately given their difference in age during the observed time-span. (left) Error*

bar showing the inferred hazard ratio coefficient and 95% CI for each regression study, as well as the sample size, n, and the number of observed events in each analysis. Note that none of the survival analyses shown are statistically significant. The complete summary for each analysis is found in Table 4.1. (right) Kaplan-Meyer survival plots for the LBC cohort stratified using 2 standard deviations of the analysed covariate.

4.3 Conclusion and 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.

Our fitness estimates are independent of the time when the mutation was acquired. In cross-sectional studies, fitness estimates are generally (inversely) correlated with the mutation rate, introducing additional uncertainty [202]. In contrast, our fitness estimates are based on the observed growth between longitudinal samples, and thus also take into account other mutations in an individual. The resulting fitness estimates are largely independent of HSC absolute numbers.

The strength of our approach, combining longitudinal data with our LiFT algorithm, is exemplified by *U2AF1* and *TP53* for which no variants were identified by a 2% VAF threshold (Figure 4.4B, 4.42C). In contrast, our LiFT method identified one *U2AF1* and 5 *TP53* variants, all of which are conferring a fitness advantage, scored as possibly damaging in our missense variant effect analysis and have been previously reported in COSMIC [259] (Figure 4.7D, Appendix 2 and 3). Moreover, we pick up the *DNMT3A R88H* variant with LiFT, but not 2% VAF thresholding, a mutation that is well reported in the context of leukaemia [80]. Therefore, for patients with those variants close clinical monitoring for early detection of disease such as leukaemia is merited.

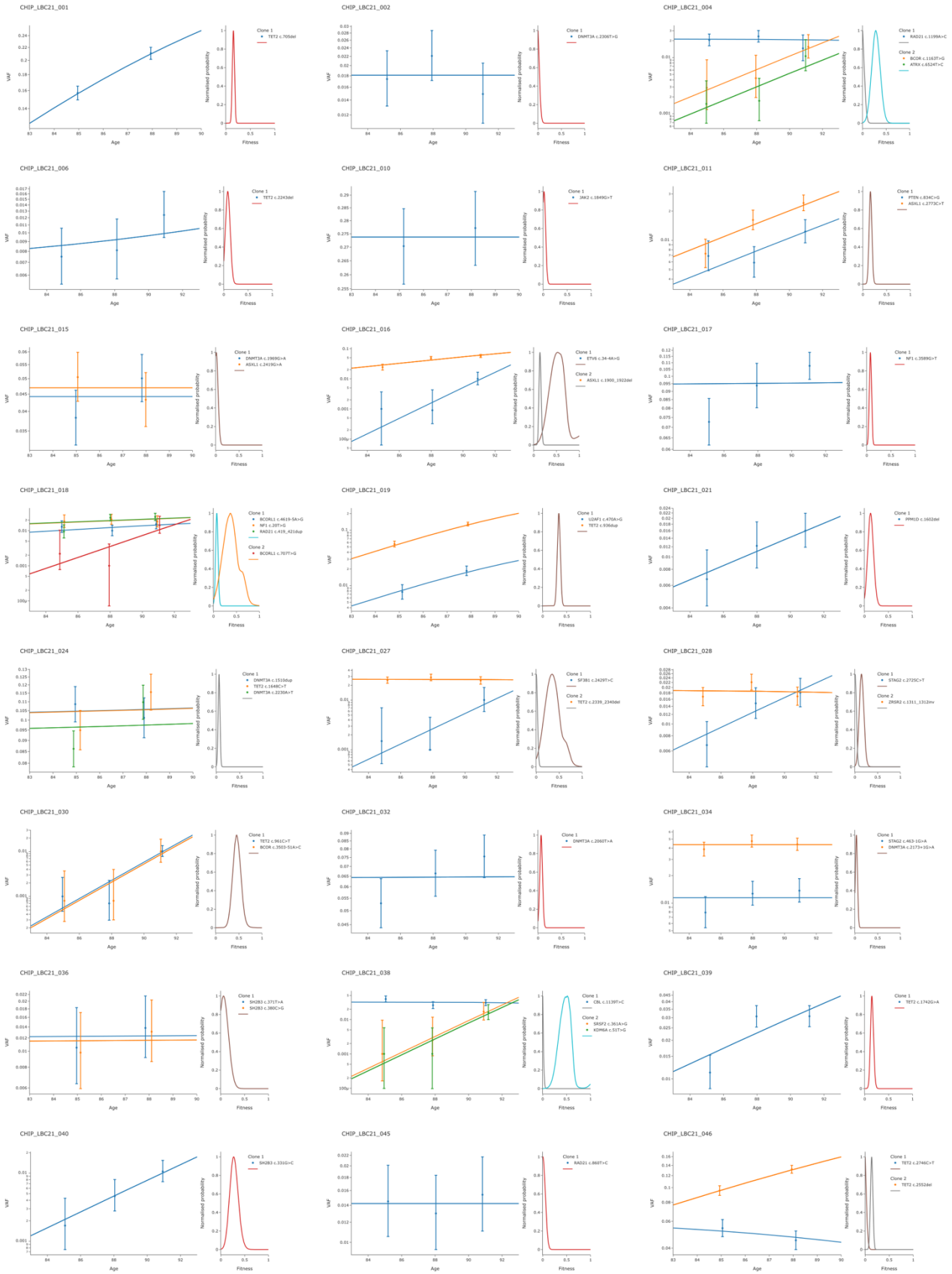


Figure 4.12: Visualisation of clonal trajectories in the LBC1921.

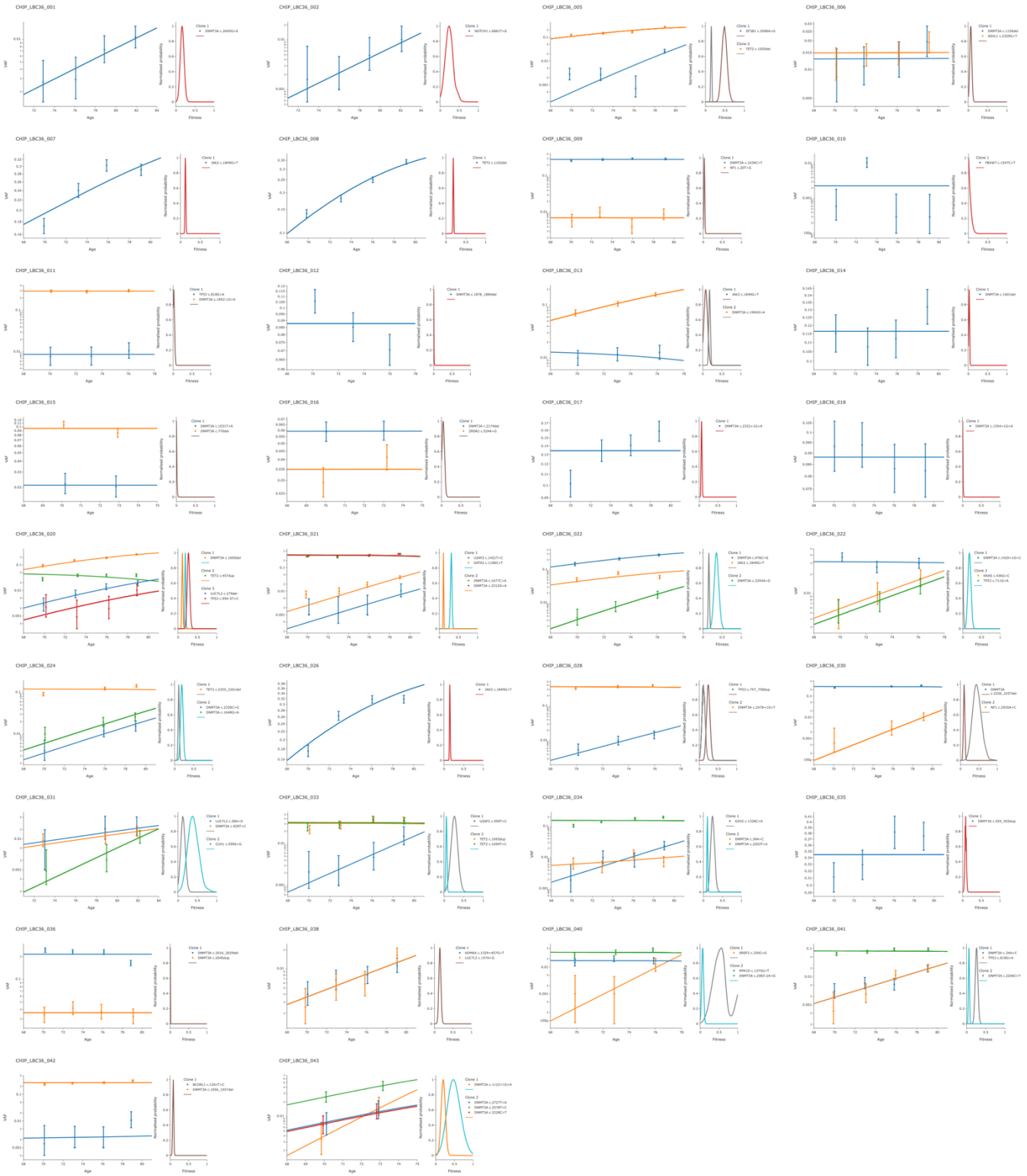


Figure 4.13: Visualisation of clonal trajectories in the LBC1936.

Combining longitudinal data with LiFT enables a personalised approach managing CHIP (Figures 4.12 and 4.13). Longitudinal data allows quantifying fitness effects even for mutations not seen in large cohorts, as cross-sectional fitness estimation requires a mutation to be observed in multiple individuals. Our method offers clinicians a way forward for patient stratification even for unique variants occurring in single individuals, since two time-points for one individual suffice to estimate fitness including uncertainty quantification. We have provided a prediction of the time required between first and second observations to be able to accurately infer fitness, depending on the initial VAF of a mutation in an individual (Figure 4.10A). For high fitness mutations (>10%) a follow-up clinical observation could be performed after only a few months, even for small clones (1% VAF or less). Conversely, the time between observations places a lower bound on the fitness that can be measured for mutations of a given VAF (Figure 4.10B). In future, these data can be used to inform time to the next appointment for close clinical monitoring of patients with clones containing highly fit variants, which will likely outcompete other clones. Using longitudinal data to better quantify and predict clonal progression in our study, however, comes with a trade-off in the lower number of participants in our cohort and limits the power of cross-sectional analysis to find associations.

In addition, our inference method aims to resolve the clonal composition of multiple mutations in an individual. Specifically, we can now infer the likely co-occurrence of mutations from longitudinal data. Current cross-sectional studies do not take into account the clonal composition of individuals and therefore make predictions of the isolated effect of a mutation. In contrast, we are able to link fitness to clones carrying a specific combination of mutations that is unique to each individual, without relying on any prior knowledge of variant-specific fitness effects (Appendix 2).

Chapter 5: Conclusions

We inevitably acquire somatic mutations in our tissues and organs as we age. The nature of this somatic mutational burden is rapidly changing how we view cancer evolution and its interactions with ageing. The discovery of clonal haematopoiesis has been an essential milestone in our understanding of the somatic evolution of cancer due to the abundance of material we have access to from purportedly healthy individuals. Work over the last decade has highlighted its age-dependent prevalence, genetic architecture and links to haematological and non-haematological disease.

Here we have characterised the somatic mutational burden in the Lothian Birth Cohort and shown a link between CH and accelerated epigenetic ageing in several published clocks. I then present one of the first longitudinal assessments of mutational fitness in CH. Here we note the importance and varying rates of clone growth between mutations in different gene contexts and propose a new method of appraising the clinical relevance of clone growth. Here I will discuss the key outcomes of this work, highlighting some of the critical questions that arise and directions for future investigation.

5.1 What is the Relevance of Accelerated Epigenetic Ageing in CH

While epigenetic age accelerations had previously been associated with a range of pathologies [220, 237–239], its links to CH had not yet been elucidated. Here we show that CH is significantly associated with biological age acceleration in several intrinsic and extrinsic clocks with effect sizes that have broadly been recapitulated in several studies since [285–287]. Why we observe accelerated epigenetic ageing in patients with CH is a matter of some debate and gets to the core of what drives mutational fitness. Two potential questions emerge from this study: 1) does CH require advanced age and an aged environment to create a positive fitness effect when they were once deleterious or neutral? Or 2) do the mutations in CH alter the profiles of their cellular outputs, creating a more inflammatory, aged environment?

Recent evidence from several functional studies has suggested that both aspects might be true. It has been shown that mutations in both TET2 and DNMT3A result in increased inflammatory outputs [84, 94, 98] and that mutated cells show increased tolerance to these environments [94, 296, 297]. There is also evidence of an interplay in the form of enhanced inflammatory crosstalk between mutated HSCs, the niche and systemic stresses [181–183]. Conversely, spliceosomal and DDR genes appear only to grow when certain conditions within the context of specific aged environments. This could allow us to use accelerated epigenetic ageing with CH status as a new proxy to assess the risk of increased clonal growth and disease outcomes [285].

CH is a substantial risk factor for numerous age-related and distal pathologies. While we were underpowered to perform an effective EWAS (epigenome-wide association study) for clonal haematopoiesis, other groups have shown substantial signatures of changing DNA methylation profiles associated with CH. The affected loci overlap regions and genes linked to cardiovascular disease and cancer progression, particularly AML [298]. When we consider this, alongside their association with epigenetic age estimates, it's compelling to think that it might be possible to create a predictive tool for CH using methylation data that captures the oligoclonal burden within an individual and their associated risk of disease progression.

To conclude, a greater functional and epidemiological study of CH will allow for a more practical understanding of the pathogenesis of this phenomenon alongside the age-dependent physiological changes that might drive it.

5.2 Why Do Different Genes Have Different Fitness Estimates

This work has highlighted the marked divergence of mutational fitness across the spectrum of common CH genes (chapter 4). This likely points to a variety of different mechanisms and the mixture of intrinsic and extrinsic factors that facilitate this growth across a lifetime. Mutations in the most common genes, DNMT3A and TET2, have relatively low fitness effects and emerge consistently throughout life and take many years to reach dominance in the haematopoietic system. Mutations in spliceosomal

genes, however, can present with large fitness estimates – even growing at rates above 50% per year - but only emerge in later life [5, 52, 101]. Our fitness scores explain, in part, why we see the patterns and prevalence of specific gene mutations in CH at a population level while also hinting at the possible mechanisms of their growth.

Somatic evolution highlights the cooperation required between mutations to achieve increased fitness within an environment. To the best of our knowledge, this platform is the first to not only look at a gene level but it can also achieve estimates of fitness at a clonal level with complex mutational contexts.

While our understanding of the variety of mechanisms that drive clone growth is in its early days, a greater understanding of both the age dependence and mechanisms of mutational fitness will be vital in enabling accurate predictions of clonal expansions and their associated disease risk. Furthermore, elucidating these mechanisms may eventually lead to targets that will allow us to slow the growth of potentially pathogenic clones within a clinical setting.

5.3 Potential Clinical Implications of Gene Fitness Estimates

CH has well-described links to haematological and non-haematological diseases [3, 199]. Translation of this knowledge to the clinical setting has thus far been complicated as it has been difficult to calculate the risk associated with complex mutational contexts at an individual level and how quickly they might grow to dominate the haematopoietic system. While our work has highlighted how specific driver mutations have well preserved fitnesses across a small population, it has also shown that fitness estimates themselves may prove to be an important metric in predicting disease risk and death.

The classical cross-sectional view of CH can provide an accurate estimate of the clonal burden in individuals and to which the risk of progression to AML has been linked [3]. However, this simple snapshot and use of arbitrary thresholds fails to account for the dynamics of CH and the presence of low VAF mutations that may still have high rates of growth. The relative stability of clone growth in genes - across this small cohort and others [101] – means that we have calculated the fitness estimate for

a mutation, we can make accurate predictions of its behaviour over time (Figure 4.10). This will allow for accurate monitoring of patient risk over long time scales.

We also present the use of fitness estimates as a new measure of disease risk. We have observed that the growth speed (fitness x VAF) of a mutation has proven to be a better predictor of all-cause mortality than clone size alone. While underpowered, we hope the expansion of our longitudinal cohort alongside projecting fitness estimates within the UK Bio Bank will help us to fully understand the potential of this method to stratify patient risk.

5.4 Final Remarks

This work adds context and perspective on CH and provides a framework for the prediction of clone growth and its associated risks. While our knowledge of CH has increased dramatically in recent times, the next phase of research must act to flesh out a more complete set of known CH drivers. The work of Jaiswal et al., and Genovese et al. described the most prevalent gene-variants, now a new generation and scale of cohort work will be needed to discover new somatic drivers and provide a better understanding of how they interact with the germline.

Alongside this, better model systems are required to allow for a more comprehensive study of the mechanisms of clonal expansion and how it interacts with the aged environment. The last ten years have transformed our view of the somatic evolution of cancer. There is no doubt the next decade will have massive implications in our understanding of CH, leading to improvements in our ability to predict, prevent and treat haematological and non-haematological disease.

Bibliography

1. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P. V., Mar, B. G., Lindsley, R. C., Mermel, C. H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.

2. 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.

3. 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.

4. 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*, R786–R787.

5. Robertson, N. A., Latorre-Crespo, E., Terradas-Terradas, M., Lemos-Portela, J., Purcell, A. C., Livesey, B. J., Hillary, R. F., Murphy, L., Fawkes, A., MacGillivray, L., et al. (2022). Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. Nat. Med. *28*, 1439–1446.

6. Doulatov, S., Notta, F., Laurenti, E., and Dick, J. E. (2012). Hematopoiesis: a human perspective. Cell Stem Cell *10*, 120–136.

7. Orkin, S. H., and Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell *132*, 631–644.

8. 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.

9. 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*, 473–478.

10. Tavassoli, M. (1991). Embryonic and fetal hemopoiesis: an overview. Blood Cells *17*, 269–81; discussion 282.

11. Rieger, M. A., and Schroeder, T. (2012). Hematopoiesis. Cold Spring Harb. Perspect. Biol. *4*.

12. Garg, S., Madkaikar, M., and Ghosh, K. (2013). Investigating cell surface markers on normal hematopoietic stem cells in three different niche conditions. Int J Stem Cells *6*, 129–133.

13. Alsinet, C., Primo, M. N., Lorenzi, V., Bello, E., Kelava, I., Jones, C. P., Vilarrasa-Blasi, R., Sancho-Serra, C., Knights, A. J., Park, J.-E., et al. (2022). Robust temporal map of

human in vitro myelopoiesis using single-cell genomics. Nat. Commun. *13*, 2885.

14. Pucella, J. N., Upadhaya, S., and Reizis, B. (2020). The Source and Dynamics of Adult Hematopoiesis: Insights from Lineage Tracing. Annu. Rev. Cell Dev. Biol. *36*, 529–550.

15. Pinho, S., and Frenette, P. S. (2019). Haematopoietic stem cell activity and interactions with the niche. Nat. Rev. Mol. Cell Biol. *20*, 303–320.

16. Pellin, D., Loperfido, M., Baricordi, C., Wolock, S. L., Montepeloso, A., Weinberg, O. K., Biffi, A., Klein, A. M., and Biasco, L. (2019). A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. Nat. Commun. *10*, 2395.

17. Ranzoni, A. M., Tangherloni, A., Berest, I., Riva, S. G., Myers, B., Strzelecka, P. M., Xu, J., Panada, E., Mohorianu, I., Zaugg, J. B., et al. (2021). Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. Cell Stem Cell *28*, 472–487.e7.

18. Xie, X., Liu, M., Zhang, Y., Wang, B., Zhu, C., Wang, C., Li, Q., Huo, Y., Guo, J., Xu, C., et al. (2021). Single-cell transcriptomic landscape of human blood cells. Natl Sci Rev *8*, nwaa180.

19. Chen, H., Albergante, L., Hsu, J. Y., Lareau, C. A., Lo Bosco, G., Guan, J., Zhou, S., Gorban, A. N., Bauer, D. E., Aryee, M. J., et al. (2019). Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. Nat. Commun. *10*, 1903.

20. Watcham, S., Kucinski, I., and Gottgens, B. (2019). New insights into hematopoietic differentiation landscapes from single-cell RNA sequencing. Blood *133*, 1415–1426.

21. Till, J. E., Mcculloch, E. A., and Siminovitch, L. (1964). A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. Proc. Natl. Acad. Sci. USA *51*, 29–36.

22. Becker, A. J., Mcculloch, E. A., Siminovitch, L., and Till, J. E. (1965). The effect of differing demands for blood cell production on dna synthesis by hemopoietic colony-forming cells of mice. Blood *26*, 296–308.

23. Galloway, J. L., Wingert, R. A., Thisse, C., Thisse, B., and Zon, L. I. (2005). Loss of gata1 but not gata2 converts erythropoiesis to myelopoiesis in zebrafish embryos. Dev. Cell *8*, 109–116.

24. Dahl, R., Iyer, S. R., Owens, K. S., Cuylear, D. D., and Simon, M. C. (2007). The transcriptional repressor GFI-1 antagonizes PU.1 activity through protein-protein interaction. J. Biol. Chem. *282*, 6473–6483.

25. Starck, J., Cohet, N., Gonnet, C., Sarrazin, S., Doubeikovskaia, Z., Doubeikovski, A., Verger, A., Duterque-Coquillaud, M., and Morle, F. (2003). Functional cross-antagonism between transcription factors FLI-1 and EKLF. Mol. Cell. Biol. *23*, 1390–1402.

26. Calvi, L. M., and Link, D. C. (2015). The hematopoietic stem cell niche in homeostasis and disease. Blood *126*, 2443–2451.

27. Batsivari, A., Haltalli, M. L. R., Passaro, D., Pospori, C., Lo Celso, C., and Bonnet, D. (2020). Dynamic responses of the haematopoietic stem cell niche to diverse stresses. Nat. Cell Biol. *22*, 7–17.

28. Niccoli, T., and Partridge, L. (2012). Ageing as a risk factor for disease. Curr. Biol. *22*, R741-52.

29. de Haan, G., Nijhof, W., and Van Zant, G. (1997). Mouse strain-dependent changes in frequency and proliferation of hematopoietic stem cells during aging: correlation between lifespan and cycling activity. Blood *89*, 1543–1550.

30. de Haan, G., and Lazare, S. S. (2018). Aging of hematopoietic stem cells. Blood *131*, 479–487.

31. Sudo, K., Ema, H., Morita, Y., and Nakauchi, H. (2000). Age-associated characteristics of murine hematopoietic stem cells. J. Exp. Med. *192*, 1273–1280.

32. Geiger, H., de Haan, G., and Florian, M. C. (2013). The ageing haematopoietic stem cell compartment. Nat. Rev. Immunol. *13*, 376–389.

33. Florian, M. C., Dörr, K., Niebel, A., Daria, D., Schrezenmeier, H., Rojewski, M., Filippi, M.-D., Hasenberg, A., Gunzer, M., Scharffetter-Kochanek, K., et al. (2012). Cdc42 activity regulates hematopoietic stem cell aging and rejuvenation. Cell Stem Cell *10*, 520–530.

34. Challen, G. A., Boles, N. C., Chambers, S. M., and Goodell, M. A. (2010). Distinct hematopoietic stem cell subtypes are differentially regulated by TGF-beta1. Cell Stem Cell *6*, 265–278.

35. Konturek-Ciesla, A., Dhapola, P., Zhang, Q., Säwén, P., Wan, H., Karlsson, G., and Bryder, D. (2023). Temporal multimodal single-cell profiling of native hematopoiesis illuminates altered differentiation trajectories with age. Cell Rep. *42*, 112304.

36. Chambers, S. M., and Goodell, M. A. (2007). Hematopoietic stem cell aging: wrinkles in stem cell potential. Stem Cell Rev *3*, 201–211.

37. Young, K., Eudy, E., Bell, R., Loberg, M. A., Stearns, T., Sharma, D., Velten, L., Haas, S., Filippi, M.-D., and Trowbridge, J. J. (2021). Decline in IGF1 in the bone marrow microenvironment initiates hematopoietic stem cell aging. Cell Stem Cell *28*, 1473–1482.e7.

38. Ergen, A. V., Boles, N. C., and Goodell, M. A. (2012). Rantes/Ccl5 influences hematopoietic stem cell subtypes and causes myeloid skewing. Blood *119*, 2500–2509.

39. Jaiswal, S., and Libby, P. (2020). Clonal haematopoiesis: connecting ageing and inflammation in cardiovascular disease. Nat. Rev. Cardiol. *17*, 137–144.

40. Busque, L., Mio, R., Mattioli, J., Brais, E., Blais, N., Lalonde, Y., Maragh, M., and Gilliland, D. G. (1996). Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. Blood *88*, 59–65.

41. Gale, R. E., Wheadon, H., and Linch, D. C. (1991). X-chromosome inactivation patterns using HPRT and PGK polymorphisms in haematologically normal and post-chemotherapy females. Br. J. Haematol. *79*, 193–197.

42. Fey, M. F., Liechti-Gallati, S., von Rohr, A., Borisch, B., Theilkäs, L., Schneider, V., Oestreicher, M., Nagel, S., Ziemiecki, A., and Tobler, A. (1994). Clonality and X-inactivation patterns in hematopoietic cell populations detected by the highly informative M27 beta DNA probe. Blood *83*, 931–938.

43. Busque, L., Paquette, Y., Provost, S., Roy, D.-C., Levine, R. L., Mollica, L., and Gilliland, D. G. (2009). Skewing of X-inactivation ratios in blood cells of aging women is confirmed by independent methodologies. Blood *113*, 3472–3474.

44. Busque, L., Patel, J. P., Figueroa, M. E., Vasanthakumar, A., Provost, S., Hamilou, Z., Mollica, L., Li, J., Viale, A., Heguy, A., et al. (2012). Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. Nat. Genet. *44*, 1179–1181.

45. Vogelstein, B., and Kinzler, K. W. (1993). The multistep nature of cancer. Trends Genet. *9*, 138–141.

46. Welch, J. S., Ley, T. J., Link, D. C., Miller, C. A., Larson, D. E., Koboldt, D. C., Wartman, L. D., Lamprecht, T. L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264–278.

47. Reya, T., Morrison, S. J., Clarke, M. F., and Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111.

48. Delhommeau, F., Dupont, S., Della Valle, V., James, C., Trannoy, S., Massé, A., Kosmider, O., Le Couedic, J.-P., Robert, F., Alberdi, A., et al. (2009). Mutation in TET2 in myeloid cancers. N. Engl. J. Med. *360*, 2289–2301.

49. Shlush, L. I., Zandi, S., Mitchell, A., Chen, W. C., Brandwein, J. M., Gupta, V., Kennedy, J. A., Schimmer, A. D., Schuh, A. C., Yee, K. W., et al. (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature *506*, 328–333.

50. Jan, M., Snyder, T. M., Corces-Zimmerman, M. R., Vyas, P., Weissman, I. L., Quake, S. R., and Majeti, R. (2012). Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. Sci. Transl. Med. *4*, 149ra118.

51. Xie, M., Lu, C., Wang, J., McLellan, M. D., Johnson, K. J., Wendl, M. C., McMichael, J. F., Schmidt, H. K., Yellapantula, V., Miller, C. A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. *20*, 1472–1478.

52. McKerrell, T., Park, N., Moreno, T., Grove, C. S., Ponstingl, H., Stephens, J., Understanding Society Scientific Group, 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.

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

54. 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.

55. Coombs, C. C., Zehir, A., Devlin, S. M., Kishtagari, A., Syed, A., Jonsson, P., Hyman,

D. M., Solit, D. B., Robson, M. E., Baselga, J., et al. (2017). Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. Cell Stem Cell *21*, 374–382.e4.

56. Bolton, K. L., Ptashkin, R. N., Gao, T., Braunstein, L., Devlin, S. M., Kelly, D., Patel, M., Berthon, A., Syed, A., Yabe, M., et al. (2020). Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nat. Genet. *52*, 1219–1226.

57. McKerrell, T., and Vassiliou, G. S. (2015). Aging as a driver of leukemogenesis. Sci. Transl. Med. *7*, 306fs38.

58. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

59. Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101.

60. Petljak, M., Alexandrov, L. B., Brammeld, J. S., Price, S., Wedge, D. C., Grossmann, S., Dawson, K. J., Ju, Y. S., Iorio, F., Tubio, J. M. C., et al. (2019). Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. Cell *176*, 1282–1294.e20.

61. Libby, P., Sidlow, R., Lin, A. E., Gupta, D., Jones, L. W., Moslehi, J., Zeiher, A., Jaiswal, S., Schulz, C., Blankstein, R., et al. (2019). Clonal Hematopoiesis: Crossroads of Aging, Cardiovascular Disease, and Cancer: JACC Review Topic of the Week. J. Am. Coll. Cardiol. *74*, 567–577.

62. Evans, M. A., Sano, S., and Walsh, K. (2021). Clonal haematopoiesis and cardiovascular disease: how low can you go? Eur. Heart J. *42*, 266–268.

63. Florez, M. A., Tran, B. T., Wathan, T. K., DeGregori, J., Pietras, E. M., and King, K. Y. (2022). Clonal hematopoiesis: Mutation-specific adaptation to environmental change. Cell Stem Cell *29*, 882–904.

64. Young, A. L., Challen, G. A., Birmann, B. M., and Druley, T. E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat. Commun. *7*, 12484.

65. Gao, L., Emperle, M., Guo, Y., Grimm, S. A., Ren, W., Adam, S., Uryu, H., Zhang, Z.-M., Chen, D., Yin, J., et al. (2020). Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. Nat. Commun. *11*, 3355.

66. Jones, P. A., and Gonzalgo, M. L. (1997). Altered DNA methylation and genome instability: a new pathway to cancer? Proc. Natl. Acad. Sci. USA *94*, 2103–2105.

67. Razin, A., and Cedar, H. (1991). DNA methylation and gene expression. Microbiol Rev *55*, 451–458.

68. Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

69. Robertson, K. D., Uzvolgyi, E., Liang, G., Talmadge, C., Sumegi, J., Gonzales, F. A., and Jones, P. A. (1999). The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. Nucleic Acids Res. *27*, 2291–2298.

70. Bröske, A.-M., Vockentanz, L., Kharazi, S., Huska, M. R., Mancini, E., Scheller, M., Kuhl, C., Enns, A., Prinz, M., Jaenisch, R., et al. (2009). DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. Nat. Genet. *41*, 1207–1215.

71. Tadokoro, Y., Ema, H., Okano, M., Li, E., and Nakauchi, H. (2007). De novo DNA methyltransferase is essential for self-renewal, but not for differentiation, in hematopoietic stem cells. J. Exp. Med. *204*, 715–722.

72. Challen, G. A., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J. S., Bock, C., Vasanthakumar, A., Gu, H., Xi, Y., et al. (2011). Dnmt3a is essential for hematopoietic stem cell differentiation. Nat. Genet. *44*, 23–31.

73. Nam, A. S., Dusaj, N., Izzo, F., Murali, R., Myers, R. M., Mouhieddine, T. H., Sotelo, J., Benbarche, S., Waarts, M., Gaiti, F., et al. (2022). Single-cell multi-omics of human clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation. Nat. Genet. *54*, 1514–1526.

74. Lu, R., Wang, P., Parton, T., Zhou, Y., Chrysovergis, K., Rockowitz, S., Chen, W.-Y., Abdel-Wahab, O., Wade, P. A., Zheng, D., et al. (2016). Epigenetic Perturbations by Arg882-Mutated DNMT3A Potentiate Aberrant Stem Cell Gene-Expression Program and Acute Leukemia Development. Cancer Cell *30*, 92–107.

75. Jeong, M., Park, H. J., Celik, H., Ostrander, E. L., Reyes, J. M., Guzman, A., Rodriguez, B., Lei, Y., Lee, Y., Ding, L., et al. (2018). Loss of dnmt3a immortalizes hematopoietic stem cells in vivo. Cell Rep. *23*, 1–10.

76. Scheller, M., Ludwig, A. K., Göllner, S., Rohde, C., Krämer, S., Stäble, S., Janssen, M., Müller, J.-A., He, L., Bäumer, N., et al. (2021). Hotspot DNMT3A mutations in clonal hematopoiesis and acute myeloid leukemia sensitize cells to azacytidine via viral mimicry response. Nat. Cancer *2*, 527–544.

77. Mayle, A., Yang, L., Rodriguez, B., Zhou, T., Chang, E., Curry, C. V., Challen, G. A., Li, W., Wheeler, D., Rebel, V. I., et al. (2015). Dnmt3a loss predisposes murine hematopoietic stem cells to malignant transformation. Blood *125*, 629–638.

78. Pløen, G. G., Nederby, L., Guldberg, P., Hansen, M., Ebbesen, L. H., Jensen, U. B., Hokland, P., and Aggerholm, A. (2014). Persistence of DNMT3A mutations at long-term remission in adult patients with AML. Br. J. Haematol. *167*, 478–486.

79. Corces-Zimmerman, M. R., Hong, W.-J., Weissman, I. L., Medeiros, B. C., and Majeti, R. (2014). Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. Proc. Natl. Acad. Sci. USA *111*, 2548–2553.

80. Ley, T. J., Ding, L., Walter, M. J., McLellan, M. D., Lamprecht, T., Larson, D. E., Kandoth, C., Payton, J. E., Baty, J., Welch, J., et al. (2010). DNMT3A mutations in acute myeloid leukemia. N. Engl. J. Med. *363*, 2424–2433.

81. Cancer Genome Atlas Research Network, Ley, T. J., Miller, C., Ding, L., Raphael, B.

J., Mungall, A. J., Robertson, A. G., Hoadley, K., Triche, T. J., Laird, P. W., et al. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059–2074.

82. Bezerra, M. F., Lima, A. S., Piqué-Borràs, M.-R., Silveira, D. R., Coelho-Silva, J. L., Pereira-Martins, D. A., Weinhäuser, I., Franca-Neto, P. L., Quek, L., Corby, A., et al. (2020). Co-occurrence of DNMT3A, NPM1, FLT3 mutations identifies a subset of acute myeloid leukemia with adverse prognosis. Blood *135*, 870–875.

83. Hormaechea Agulla, D., Matatall, K. A., Le, D., Kain, B. N., Jaksik, R., Kimmel, M., Challen, G., and King, K. Y. (2019). Infection Is a Driver of Dnmt3a-Mutant Clonal Hematopoiesis. Blood *134*, 817–817.

84. Zhang, C. R. C., Nix, D., Gregory, M., Ciorba, M. A., Ostrander, E. L., Newberry, R. D., Spencer, D. H., and Challen, G. A. (2019). Inflammatory cytokines promote clonal hematopoiesis with specific mutations in ulcerative colitis patients. Exp. Hematol. *80*, 36–41.e3.

85. Hormaechea-Agulla, D., Matatall, K. A., Le, D. T., Kain, B., Long, X., Kus, P., Jaksik, R., Challen, G. A., Kimmel, M., and King, K. Y. (2021). Chronic infection drives Dnmt3a-loss-of-function clonal hematopoiesis via IFNγ signaling. Cell Stem Cell *28*, 1428–1442.e6.

86. Kunimoto, H., and Nakajima, H. (2021). TET2: A cornerstone in normal and malignant hematopoiesis. Cancer Sci. *112*, 31–40.

87. Rasmussen, K. D., and Helin, K. (2016). Role of TET enzymes in DNA methylation, development, and cancer. Genes Dev. *30*, 733–750.

88. Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930–935.

89. Xu, Y.-P., Lv, L., Liu, Y., Smith, M. D., Li, W.-C., Tan, X.-M., Cheng, M., Li, Z., Bovino, M., Aubé, J., et al. (2019). Tumor suppressor TET2 promotes cancer immunity and immunotherapy efficacy. J. Clin. Invest. *129*, 4316–4331.

90. Pan, F., Weeks, O., Yang, F.-C., and Xu, M. (2015). The TET2 interactors and their links to hematological malignancies. IUBMB Life *67*, 438–445.

91. Jiang, S. (2020). Tet2 at the interface between cancer and immunity. Commun. Biol. *3*, 667.

92. Pan, F., Wingo, T. S., Zhao, Z., Gao, R., Makishima, H., Qu, G., Lin, L., Yu, M., Ortega, J. R., Wang, J., et al. (2017). Tet2 loss leads to hypermutagenicity in haematopoietic stem/progenitor cells. Nat. Commun. *8*, 15102.

93. Moran-Crusio, K., Reavie, L., Shih, A., Abdel-Wahab, O., Ndiaye-Lobry, D., Lobry, C., Figueroa, M. E., Vasanthakumar, A., Patel, J., Zhao, X., et al. (2011). Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. Cancer Cell *20*, 11–24.

94. Zhang, Q., Zhao, K., Shen, Q., Han, Y., Gu, Y., Li, X., Zhao, D., Liu, Y., Wang, C.,

Zhang, X., et al. (2015). Tet2 is required to resolve inflammation by recruiting Hdac2 to specifically repress IL-6. Nature *525*, 389–393.

95. Ko, M., Bandukwala, H. S., An, J., Lamperti, E. D., Thompson, E. C., Hastie, R., Tsangaratou, A., Rajewsky, K., Koralov, S. B., and Rao, A. (2011). Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. Proc. Natl. Acad. Sci. USA *108*, 14566–14571.

96. Quivoron, C., Couronné, L., Della Valle, V., Lopez, C. K., Plo, I., Wagner-Ballon, O., Do Cruzeiro, M., Delhommeau, F., Arnulf, B., Stern, M.-H., et al. (2011). TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer Cell *20*, 25–38.

97. Jaiswal, S., Natarajan, P., Silver, A. J., Gibson, C. J., Bick, A. G., Shvartz, E., McConkey, M., Gupta, N., Gabriel, S., Ardissino, D., et al. (2017). Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. N. Engl. J. Med. *377*, 111–121.

98. Fuster, J. J., MacLauchlan, S., Zuriaga, M. A., Polackal, M. N., Ostriker, A. C., Chakraborty, R., Wu, C.-L., Sano, S., Muralidharan, S., Rius, C., et al. (2017). Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. Science *355*, 842–847.

99. Bick, A. G., Weinstock, J. S., Nandakumar, S. K., Fulco, C. P., Bao, E. L., Zekavat, S. M., Szeto, M. D., Liao, X., Leventhal, M. J., Nasser, J., et al. (2020). Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature *586*, 763–768.

100. Buscarlet, M., Provost, S., Zada, Y. F., Barhdadi, A., Bourgoin, V., Lépine, G., Mollica, L., Szuber, N., Dubé, M.-P., and Busque, L. (2017). DNMT3A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. Blood *130*, 753–762.

101. Fabre, M. A., de Almeida, J. G., Fiorillo, E., Mitchell, E., Damaskou, A., Rak, J., Orrù, V., Marongiu, M., Chapman, M. S., Vijayabaskar, M. S., et al. (2022). The longitudinal dynamics and natural history of clonal haematopoiesis. Nature *606*, 335–342.

102. Izzo, F., Lee, S. C., Poran, A., Chaligne, R., Gaiti, F., Gross, B., Murali, R. R., Deochand, S. D., Ang, C., Jones, P. W., et al. (2020). DNA methylation disruption reshapes the hematopoietic differentiation landscape. Nat. Genet. *52*, 378–387.

103. Zhang, X., Su, J., Jeong, M., Ko, M., Huang, Y., Park, H. J., Guzman, A., Lei, Y., Huang, Y.-H., Rao, A., et al. (2016). DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. Nat. Genet. *48*, 1014–1023.

104. Asada, S., Fujino, T., Goyama, S., and Kitamura, T. (2019). The role of ASXL1 in hematopoiesis and myeloid malignancies. Cell Mol. Life Sci. *76*, 2511–2523.

105. Fujino, T., and Kitamura, T. (2020). ASXL1 mutation in clonal hematopoiesis. Exp. Hematol. *83*, 74–84.

106. Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B.-M., Armstrong, H., Mifsud, B., Dimitrova, E., Matheson, L., Tavares-Cadete, F., Furlan-Magaril, M., et al. (2015). Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem

cell genome. Nat. Genet. *47*, 1179–1186.

107. Schwartz, Y. B., and Pirrotta, V. (2013). A new world of Polycombs: unexpected partnerships and emerging functions. Nat. Rev. Genet. *14*, 853–864.

108. Abdel-Wahab, O., Adli, M., LaFave, L. M., Gao, J., Hricik, T., Shih, A. H., Pandey, S., Patel, J. P., Chung, Y. R., Koche, R., et al. (2012). ASXL1 mutations promote myeloid transformation through loss of PRC2-mediated gene repression. Cancer Cell *22*, 180–193.

109. Dawoud, A. A. Z., Tapper, W. J., and Cross, N. C. P. (2020). Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking. Leukemia *34*, 2660–2672.

110. Avagyan, S., Henninger, J. E., Mannherz, W. P., Mistry, M., Yoon, J., Yang, S., Weber, M. C., Moore, J. L., and Zon, L. I. (2021). Resistance to inflammation underlies enhanced fitness in clonal hematopoiesis. Science *374*, 768–772.

111. Mead, A. J., and Mullally, A. (2017). Myeloproliferative neoplasm stem cells. Blood *129*, 1607–1616.

112. McLornan, D., Percy, M., and McMullin, M. F. (2006). JAK2 V617F: a single mutation in the myeloproliferative group of disorders. Ulster Med J *75*, 112–119.

113. Seif, F., Khoshmirsafa, M., Aazami, H., Mohsenzadegan, M., Sedighi, G., and Bahar, M. (2017). The role of JAK-STAT signaling pathway and its regulators in the fate of T helper cells. Cell Commun. Signal. *15*, 23.

114. Hu, X., Li, J., Fu, M., Zhao, X., and Wang, W. (2021). The JAK/STAT signaling pathway: from bench to clinic. Signal Transduct. Target. Ther. *6*, 402.

115. Hubbard, S. R. (2017). Mechanistic Insights into Regulation of JAK2 Tyrosine Kinase. Front. Endocrinol. (Lausanne) *8*, 361.

116. Perner, F., Perner, C., Ernst, T., and Heidel, F. H. (2019). Roles of JAK2 in aging, inflammation, hematopoiesis and malignant transformation. Cells *8*.

117. Misaka, T., Kimishima, Y., Yokokawa, T., Ikeda, K., and Takeishi, Y. (2022). Clonal hematopoiesis and cardiovascular diseases: role of JAK2V617F. J Cardiol.

118. James, C., Ugo, V., Le Couédic, J.-P., Staerk, J., Delhommeau, F., Lacout, C., Garçon, L., Raslova, H., Berger, R., Bennaceur-Griscelli, A., et al. (2005). A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature *434*, 1144–1148.

119. Tefferi, A. (2021). Primary myelofibrosis: 2021 update on diagnosis, risk-stratification and management. Am. J. Hematol. *96*, 145–162.

120. Kralovics, R., Passamonti, F., Buser, A. S., Teo, S.-S., Tiedt, R., Passweg, J. R., Tichelli, A., Cazzola, M., and Skoda, R. C. (2005). A gain-of-function mutation of JAK2 in myeloproliferative disorders. N. Engl. J. Med. *352*, 1779–1790.

121. McKerrell, T., Park, N., Chi, J., Collord, G., Moreno, T., Ponstingl, H., Dias, J., Gerasimou, P., Melanthiou, K., Prokopiou, C., et al. (2017). JAK2 V617F hematopoietic clones are present several years prior to MPN diagnosis and follow

different expansion kinetics. Blood Adv. *1*, 968–971.

122. Sochacki, A., Zhao, S., Bejan, C. A., Spaulding, T., Stockton, S., Silver, A., Dorand, D., Zhang, S., Stricker, T., Xu, Y., et al. (2019). JAK2V617F clonal hematopoiesis stratifies by peripheral blood counts. Blood *134*, 1203–1203.

123. Sano, S., Wang, Y., Yura, Y., Sano, M., Oshima, K., Yang, Y., Katanasaka, Y., Min, K.-D., Matsuura, S., Ravid, K., et al. (2019). JAK2V617F -Mediated Clonal Hematopoiesis Accelerates Pathological Remodeling in Murine Heart Failure. JACC Basic Transl. Sci. *4*, 684–697.

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

125. Darnell, J. E. (1978). Implications of RNA-RNA splicing in evolution of eukaryotic cells. Science *202*, 1257–1260.

126. Patel, A. A., and Steitz, J. A. (2003). Splicing double: insights from the second spliceosome. Nat. Rev. Mol. Cell Biol. *4*, 960–970.

127. Matera, A. G., and Wang, Z. (2014). A day in the life of the spliceosome. Nat. Rev. Mol. Cell Biol. *15*, 108–121.

128. Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

129. Holly, A. C., Melzer, D., Pilling, L. C., Fellows, A. C., Tanaka, T., Ferrucci, L., and Harries, L. W. (2013). Changes in splicing factor expression are associated with advancing age in man. Mech. Ageing Dev. *134*, 356–366.

130. Kahles, A., Lehmann, K.-V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Cancer Genome Atlas Research Network, et al. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell *34*, 211–224.e6.

131. Chen, L., Kostadima, M., Martens, J. H. A., Canu, G., Garcia, S. P., Turro, E., Downes, K., Macaulay, I. C., Bielczyk-Maczynska, E., Coe, S., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. Science *345*, 1251033.

132. Wong, J. J.-L., Ritchie, W., Ebner, O. A., Selbach, M., Wong, J. W. H., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., et al. (2013). Orchestrated intron retention regulates normal granulocyte differentiation. Cell *154*, 583–595.

133. Edwards, C. R., Ritchie, W., Wong, J. J.-L., Schmitz, U., Middleton, R., An, X., Mohandas, N., Rasko, J. E. J., and Blobel, G. A. (2016). A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. Blood *127*, e24–e34.

134. Walter, M. J., and Graubert, T. A. (2016). Clinical implications of spliceosome mutations: epidemiology, clonal hematopoiesis, and potential therapeutic strategies. Blood *128*, SCI-19-SCI-19.

135. Nagata, Y., Makishima, H., Kerr, C. M., Przychodzen, B. P., Aly, M., Goyal, A., Awada,

H., Asad, M. F., Kuzmanovic, T., Suzuki, H., et al. (2019). Invariant patterns of clonal succession determine specific clinical features of myelodysplastic syndromes. Nat. Commun. *10*, 5386.

136. Seiler, M., Peng, S., Agrawal, A. A., Palacino, J., Teng, T., Zhu, P., Smith, P. G., Cancer Genome Atlas Research Network, Buonamici, S., and Yu, L. (2018). Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. Cell Rep. *23*, 282–296.e4.

137. Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature *478*, 64–69.

138. Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., Bowen, D., Pellagatti, A., Wainscoat, J. S., Hellstrom-Lindberg, E., Gambacorti-Passerini, C., et al. (2011). Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N. Engl. J. Med. *365*, 1384–1395.

139. Gaiti, F., Chamely, P., Hawkins, A. G., Cortes-Lopez, M., Swett, A. D., Ganesan, S., Mouhieddine, T. H., Dai, X., Kluegel, L., Chen, C., et al. (2022). Single-cell multi-omics defines the cell-type specific impact of splicing aberrations in human hematopoietic clonal outgrowths. BioRxiv.

140. Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., Schnittger, S., Sanada, M., Kon, A., Alpermann, T., et al. (2014). Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. Leukemia *28*, 241–247.

141. Wu, S.-J., Kuo, Y.-Y., Hou, H.-A., Li, L.-Y., Tseng, M.-H., Huang, C.-F., Lee, F.-Y., Liu, M.-C., Liu, C.-W., Lin, C.-T., et al. (2012). The clinical implication of SRSF2 mutation in patients with myelodysplastic syndrome and its stability during disease evolution. Blood *120*, 3106–3111.

142. Mantovani, F., Collavin, L., and Del Sal, G. (2019). Mutant p53 as a guardian of the cancer cell. Cell Death Differ. *26*, 199–212.

143. Hafner, A., Bulyk, M. L., Jambhekar, A., and Lahav, G. (2019). The multiple mechanisms that regulate p53 activity and cell fate. Nat. Rev. Mol. Cell Biol. *20*, 199–210.

144. Mijit, M., Caracciolo, V., Melillo, A., Amicarelli, F., and Giordano, A. (2020). Role of p53 in the regulation of cellular senescence. Biomolecules *10*.

145. Aubrey, B. J., Kelly, G. L., Janic, A., Herold, M. J., and Strasser, A. (2018). How does p53 induce apoptosis and how does this relate to p53-mediated tumour suppression? Cell Death Differ. *25*, 104–113.

146. Bondar, T., and Medzhitov, R. (2010). p53-mediated hematopoietic stem and progenitor cell competition. Cell Stem Cell *6*, 309–322.

147. Dumble, M., Moore, L., Chambers, S. M., Geiger, H., Van Zant, G., Goodell, M. A., and Donehower, L. A. (2007). The impact of altered p53 dosage on hematopoietic stem cell dynamics during aging. Blood *109*, 1736–1742.

148. Bowman, R. L., Busque, L., and Levine, R. L. (2018). Clonal hematopoiesis and

evolution to hematopoietic malignancies. Cell Stem Cell *22*, 157–170.

149. Hsu, J. I., Dayaram, T., Tovy, A., De Braekeleer, E., Jeong, M., Wang, F., Zhang, J., Heffernan, T. P., Gera, S., Kovacs, J. J., et al. (2018). PPM1D mutations drive clonal hematopoiesis in response to cytotoxic chemotherapy. Cell Stem Cell *23*, 700–713.e6.

150. Viny, A. D., Ott, C. J., Spitzer, B., Rivas, M., Meydan, C., Papalexi, E., Yelin, D., Shank, K., Reyes, J., Chiu, A., et al. (2015). Dose-dependent role of the cohesin complex in normal and malignant hematopoiesis. J. Exp. Med. *212*, 1819–1832.

151. Mazumdar, C., and Majeti, R. (2017). The role of mutations in the cohesin complex in acute myeloid leukemia. Int J Hematol *105*, 31–36.

152. Merkenschlager, M., and Odom, D. T. (2013). CTCF and cohesin: linking gene regulatory elements with their targets. Cell *152*, 1285–1297.

153. Qi, Q., Cheng, L., Tang, X., He, Y., Li, Y., Yee, T., Shrestha, D., Feng, R., Xu, P., Zhou, X., et al. (2021). Dynamic CTCF binding directly mediates interactions among cis-regulatory elements essential for hematopoiesis. Blood *137*, 1327–1339.

154. Mill, C. P., Fiskus, W., DiNardo, C. D., Birdwell, C., Davis, J. A., Kadia, T. M., Takahashi, K., Short, N., Daver, N., Ohanian, M., et al. (2022). Effective therapy for AML with RUNX1 mutation by cotreatment with inhibitors of protein translation and BCL2. Blood *139*, 907–921.

155. Mitchell, E., Spencer Chapman, M., Williams, N., Dawson, K., Mende, N., Calderbank, E. F., Jung, H., Mitchell, T. J., Coorens, T., Spencer, D. H., et al. (2021). Clonal dynamics of haematopoiesis across the human lifespan. BioRxiv.

156. Kar, S. P., Quiros, P. M., Gu, M., Jiang, T., Mitchell, J., Langdon, R., Iyer, V., Barcena, C., Vijayabaskar, M. S., Fabre, M. A., et al. (2022). Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. Nat. Genet. *54*, 1155–1166.

157. Sano, S., Horitani, K., Ogawa, H., Halvardson, J., Chavkin, N. W., Wang, Y., Sano, M., Mattisson, J., Hata, A., Danielsson, M., et al. (2022). Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. Science *377*, 292–297.

158. Forsberg, L. A., Rasi, C., Malmqvist, N., Davies, H., Pasupulati, S., Pakalapati, G., Sandgren, J., Diaz de Ståhl, T., Zaghlool, A., Giedraitis, V., et al. (2014). Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nat. Genet. *46*, 624–628.

159. Zekavat, S. M., Lin, S.-H., Bick, A. G., Liu, A., Paruchuri, K., Wang, C., Uddin, M. M., Ye, Y., Yu, Z., Liu, X., et al. (2021). Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. Nat. Med. *27*, 1012–1024.

160. Levin, M. G., Nakao, T., Zekavat, S. M., Koyama, S., Bick, A. G., Niroula, A., Ebert, B., Damrauer, S. M., and Natarajan, P. (2022). Genetics of smoking and risk of clonal hematopoiesis. Sci. Rep. *12*, 7248.

161. Kessler, M. D., Damask, A., O'Keeffe, S., Michael Van Meter, M., Banerjee, N., Semrau, S., Li, D., Watanabe, K., Horowitz, J., Houvras, Y., et al. (2022). Exome

sequencing of 628,388 individuals identifies common and rare variant associations with clonal hematopoiesis phenotypes. medRxiv.

162. Shay, J. W., and Wright, W. E. (2019). Telomeres and telomerase: three decades of progress. Nat. Rev. Genet. *20*, 299–309.

163. Demanelis, K., Jasmine, F., Chen, L. S., Chernoff, M., Tong, L., Delgado, D., Zhang, C., Shinkle, J., Sabarinathan, M., Lin, H., et al. (2020). Determinants of telomere length across human tissues. Science *369*.

164. Silver, A. J., Bick, A. G., and Savona, M. R. (2021). Germline risk of clonal haematopoiesis. Nat. Rev. Genet. *22*, 603–617.

165. Brown, D. W., Cato, L. D., Zhao, Y., Nandakumar, S. K., Bao, E. L., Rehling, T., Song, L., Yu, K., Chanock, S. J., Perry, J. R. B., et al. (2022). Shared and distinct genetic etiologies for different types of clonal hematopoiesis. BioRxiv.

166. Weinstock, J. S., Gopakumar, J., Burugula, B. B., Uddin, M. M., Jahn, N., Belk, J. A., Daniel, B., Ly, N., Mack, T. M., Laurie, C. A., et al. (2021). Clonal hematopoiesis is driven by aberrant activation of TCL1A. BioRxiv.

167. SanMiguel, J. M., Young, K., and Trowbridge, J. J. (2020). Hand in hand: intrinsic and extrinsic drivers of aging and clonal hematopoiesis. Exp. Hematol. *91*, 1–9.

168. Tikhonova, A. N., Dolgalev, I., Hu, H., Sivaraj, K. K., Hoxha, E., Cuesta-Domínguez, Á., Pinho, S., Akhmetzyanova, I., Gao, J., Witkowski, M., et al. (2019). The bone marrow microenvironment at single-cell resolution. Nature *569*, 222–228.

169. Baryawno, N., Przybylski, D., Kowalczyk, M. S., Kfoury, Y., Severe, N., Gustafsson, K., Kokkaliaris, K. D., Mercier, F., Tabaka, M., Hofree, M., et al. (2019). A cellular taxonomy of the bone marrow stroma in homeostasis and leukemia. Cell *177*, 1915–1932.e16.

170. Baccin, C., Al-Sabah, J., Velten, L., Helbling, P. M., Grünschläger, F., Hernández-Malmierca, P., Nombela-Arrieta, C., Steinmetz, L. M., Trumpp, A., and Haas, S. (2020). Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nat. Cell Biol. *22*, 38–48.

171. Saçma, M., Pospiech, J., Bogeska, R., de Back, W., Mallm, J.-P., Sakk, V., Soller, K., Marka, G., Vollmer, A., Karns, R., et al. (2019). Haematopoietic stem cells in perisinusoidal niches are protected from ageing. Nat. Cell Biol. *21*, 1309–1320.

172. Frick, M., Chan, W., Arends, C. M., Hablesreiter, R., Halik, A., Heuser, M., Michonneau, D., Blau, O., Hoyer, K., Christen, F., et al. (2019). Role of Donor Clonal Hematopoiesis in Allogeneic Hematopoietic Stem-Cell Transplantation. J. Clin. Oncol. *37*, 375–385.

173. Wong, W. H., Bhatt, S., Trinkaus, K., Pusic, I., Elliott, K., Mahajan, N., Wan, F., Switzer, G. E., Confer, D. L., DiPersio, J., et al. (2020). Engraftment of rare, pathogenic donor hematopoietic mutations in unrelated hematopoietic stem cell transplantation. Sci. Transl. Med. *12*.

174. Boettcher, S., Wilk, C. M., Singer, J., Beier, F., Burcklen, E., Beisel, C., Ventura Ferreira, M. S., Gourri, E., Gassner, C., Frey, B. M., et al. (2020). Clonal

hematopoiesis in donors and long-term survivors of related allogeneic hematopoietic stem cell transplantation. Blood *135*, 1548–1559.

175. Gibson, C. J., Kim, H. T., Zhao, L., Murdock, H. M., Hambley, B., Ogata, A., Madero-Marroquin, R., Wang, S., Green, L., Fleharty, M., et al. (2022). Donor clonal hematopoiesis and recipient outcomes after transplantation. J. Clin. Oncol. *40*, 189–201.

176. Bolton, K. L., Ptashkin, R. N., Gao, T., Braunstein, L., Devlin, S. M., Kelly, D., Patel, M., Berthon, A., Syed, A., Yabe, M., et al. (2019). Oncologic therapy shapes the fitness landscape of clonal hematopoiesis. BioRxiv.

177. Kovtonyuk, L. V., Fritsch, K., Feng, X., Manz, M. G., and Takizawa, H. (2016). Inflamm-Aging of Hematopoiesis, Hematopoietic Stem Cells, and the Bone Marrow Microenvironment. Front. Immunol. *7*, 502.

178. Yamashita, M., and Passegué, E. (2019). TNF-α Coordinates Hematopoietic Stem Cell Survival and Myeloid Regeneration. Cell Stem Cell *25*, 357–372.e7.

179. Frisch, B. J., Hoffman, C. M., Latchney, S. E., LaMere, M. W., Myers, J., Ashton, J., Li, A. J., Saunders, J., Palis, J., Perkins, A. S., et al. (2019). Aged marrow macrophages expand platelet-biased hematopoietic stem cells via interleukin-1B. JCI Insight.

180. Cai, Z., Kotzin, J. J., Ramdas, B., Chen, S., Nelanuthala, S., Palam, L. R., Pandey, R., Mali, R. S., Liu, Y., Kelley, M. R., et al. (2018). Inhibition of Inflammatory Signaling in Tet2 Mutant Preleukemic Cells Mitigates Stress-Induced Abnormalities and Clonal Hematopoiesis. Cell Stem Cell *23*, 833–849.e5.

181. Cook, E. K., Izukawa, T., Young, S., Rosen, G., Jamali, M., Zhang, L., Johnson, D., Bain, E., Hilland, J., Ferrone, C. K., et al. (2019). Comorbid and inflammatory characteristics of genetic subtypes of clonal hematopoiesis. Blood Adv. *3*, 2482–2486.

182. Baldridge, M. T., King, K. Y., Boles, N. C., Weksberg, D. C., and Goodell, M. A. (2010). Quiescent haematopoietic stem cells are activated by IFN-gamma in response to chronic infection. Nature *465*, 793–797.

183. Trowbridge, J. J., and Starczynowski, D. T. (2021). Innate immune pathways and inflammation in hematopoietic aging, clonal hematopoiesis, and MDS. J. Exp. Med. *218*.

184. Ho, T. T., Dellorusso, P. V., Verovskaya, E. V., Bakker, S. T., Flach, J., Smith, L. K., Ventura, P. B., Lansinger, O. M., Hérault, A., Zhang, S. Y., et al. (2021). Aged hematopoietic stem cells are refractory to bloodborne systemic rejuvenation interventions. J. Exp. Med. *218*.

185. Bewersdorf, J. P., Ardasheva, A., Podoltsev, N. A., Singh, A., Biancon, G., Halene, S., and Zeidan, A. M. (2019). From clonal hematopoiesis to myeloid leukemia and what happens in between: Will improved understanding lead to new therapeutic and preventive opportunities? Blood Rev *37*, 100587.

186. Vetrie, D., Helgason, G. V., and Copland, M. (2020). The leukaemia stem cell: similarities, differences and clinical prospects in CML and AML. Nat. Rev. Cancer *20*, 158–173.

187. Ortmann, C. A., Kent, D. G., Nangalia, J., Silber, Y., Wedge, D. C., Grinfeld, J., Baxter, E. J., Massie, C. E., Papaemmanuil, E., Menon, S., et al. (2015). Effect of mutation order on myeloproliferative neoplasms. N. Engl. J. Med. *372*, 601–612.

188. Kent, D. G., and Green, A. R. (2017). Order matters: the order of somatic mutations influences cancer evolution. Cold Spring Harb. Perspect. Med. *7*.

189. Greenfield, G., McMullin, M. F., and Mills, K. (2021). Molecular pathogenesis of the myeloproliferative neoplasms. J Hematol Oncol *14*, 103.

190. Desai, P., Mencia-Trinchant, N., Savenkov, O., Simon, M. S., Cheang, G., Lee, S., Samuel, M., Ritchie, E. K., Guzman, M. L., Ballman, K. V., et al. (2018). Somatic mutations precede acute myeloid leukemia years before diagnosis. Nat. Med. *24*, 1015–1023.

191. 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). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science *348*, 880–886.

192. Jaiswal, S. (2020). Clonal hematopoiesis and nonhematologic disorders. Blood *136*, 1606–1614.

193. Yu, B., Roberts, M. B., Raffield, L. M., Zekavat, S. M., Nguyen, N. Q. H., Biggs, M. L., Brown, M. R., Griffin, G., Desai, P., Correa, A., et al. (2021). Supplemental association of clonal hematopoiesis with incident heart failure. J. Am. Coll. Cardiol. *78*, 42–52.

194. Dorsheimer, L., Assmus, B., Rasper, T., Ortmann, C. A., Ecke, A., Abou-El-Ardat, K., Schmid, T., Brüne, B., Wagner, S., Serve, H., et al. (2019). Association of mutations contributing to clonal hematopoiesis with prognosis in chronic ischemic heart failure. JAMA Cardiol. *4*, 25–33.

195. Sano, S., Oshima, K., Wang, Y., MacLauchlan, S., Katanasaka, Y., Sano, M., Zuriaga, M. A., Yoshiyama, M., Goukassian, D., Cooper, M. A., et al. (2018). Tet2-Mediated Clonal Hematopoiesis Accelerates Heart Failure Through a Mechanism Involving the IL-1β/NLRP3 Inflammasome. J. Am. Coll. Cardiol. *71*, 875–886.

196. Sano, S., Oshima, K., Wang, Y., Katanasaka, Y., Sano, M., and Walsh, K. (2018). CRISPR-Mediated Gene Editing to Assess the Roles of Tet2 and Dnmt3a in Clonal Hematopoiesis and Cardiovascular Disease. Circ. Res. *123*, 335–341.

197. Bick, A. G., Pirruccello, J. P., Griffin, G. K., Gupta, N., Gabriel, S., Saleheen, D., Libby, P., Kathiresan, S., and Natarajan, P. (2020). Genetic interleukin 6 signaling deficiency attenuates cardiovascular risk in clonal hematopoiesis. Circulation *141*, 124–131.

198. Vlasschaert, C., Heimlich, J. B., Rauh, M. J., Natarajan, P., and Bick, A. G. (2023). Interleukin-6 Receptor Polymorphism Attenuates Clonal Hematopoiesis-Mediated Coronary Artery Disease Risk Among 451 180 Individuals in the UK Biobank. Circulation *147*, 358–360.

199. Calvillo-Argüelles, O., Jaiswal, S., Shlush, L. I., Moslehi, J. J., Schimmer, A., Barac, A., and Thavendiranathan, P. (2019). Connections between clonal hematopoiesis, cardiovascular disease, and cancer: A review. JAMA Cardiol. *4*, 380–387.

200. Steensma, D. P., and Bolton, K. L. (2020). What to tell your patient with clonal hematopoiesis and why: insights from 2 specialized clinics. Blood *136*, 1623–1631.

201. van Zeventer, I. A., de Graaf, A. O., Wouters, H. J. C. M., van der Reijden, B. A., van der Klauw, M. M., de Witte, T., Jonker, M. A., Malcovati, L., Jansen, J. H., and Huls, G. (2020). Mutational spectrum and dynamics of clonal hematopoiesis in anemia of older individuals. Blood *135*, 1161–1170.

202. 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 *367*, 1449–1454.

203. Poon, Y. P. G., Watson, C. J., Fisher, D. S., and Blundell, J. R. (2020). Synonymous mutations reveal genome-wide driver mutation rates in healthy tissues. BioRxiv.

204. Field, A. E., Robertson, N. A., Wang, T., Havas, A., Ideker, T., and Adams, P. D. (2018). DNA methylation clocks in aging: categories, causes, and consequences. Mol. Cell *71*, 882–895.

205. Jansz, N. (2019). DNA methylation dynamics at transposable elements in mammals. Essays Biochem *63*, 677–689.

206. Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

207. Moore, L. D., Le, T., and Fan, G. (2013). DNA Methylation and Its Basic Function. Neuropsychopharmacology *38*, 23–38.

208. Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321–326.

209. Chédin, F. (2011). The DNMT3 family of mammalian de novo DNA methyltransferases. Prog Mol Biol Transl Sci *101*, 255–285.

210. Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., and Schübeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. *9*, e1003994.

211. Rulands, S., Lee, H. J., Clark, S. J., Angermueller, C., Smallwood, S. A., Krueger, F., Mohammed, H., Dean, W., Nichols, J., Rugg-Gunn, P., et al. (2018). Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency. Cell Syst. *7*, 63–76.e12.

212. Unnikrishnan, A., Hadad, N., Masser, D. R., Jackson, J., Freeman, W. M., and Richardson, A. (2018). Revisiting the genomic hypomethylation hypothesis of aging. Ann. N. Y. Acad. Sci. *1418*, 69–79.

213. Berger, S. L., and Sassone-Corsi, P. (2016). Metabolic signaling to chromatin. Cold Spring Harb. Perspect. Biol. *8*.

214. Issa, J.-P. (2014). Aging and epigenetic drift: a vicious cycle. J. Clin. Invest. *124*, 24–29.

215. Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., Heine-Suñer, D., Cigudosa, J. C., Urioste, M., Benitez, J., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc. Natl. Acad. Sci. USA

102, 10604–10609.

216. Teschendorff, A. E., Menon, U., Gentry-Maharaj, A., Ramus, S. J., Weisenberger, D. J., Shen, H., Campan, M., Noushmehr, H., Bell, C. G., Maxwell, A. P., et al. (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. *20*, 440–446.

217. Cole, J. J., Robertson, N. A., Rather, M. I., Thomson, J. P., McBryan, T., Sproul, D., Wang, T., Brock, C., Clark, W., Ideker, T., et al. (2017). Diverse interventions that extend mouse lifespan suppress shared age-associated epigenetic changes at critical gene regulatory regions. Genome Biol. *18*, 58.

218. Wang, T., Tsui, B., Kreisberg, J. F., Robertson, N. A., Gross, A. M., Yu, M. K., Carter, H., Brown-Borg, H. M., Adams, P. D., and Ideker, T. (2017). Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. Genome Biol. *18*, 57.

219. Fraga, M. F., and Esteller, M. (2007). Epigenetics and aging: the targets and the marks. Trends Genet. *23*, 413–418.

220. Oblak, L., van der Zaag, J., Higgins-Chen, A. T., Levine, M. E., and Boks, M. P. (2021). A systematic review of biological, social and environmental factors associated with epigenetic clock acceleration. Ageing Res Rev *69*, 101348.

221. Jylhävä, J., Pedersen, N. L., and Hägg, S. (2017). Biological age predictors. EBioMedicine *21*, 29–36.

222. Simpson, D. J., and Chandra, T. (2021). Epigenetic age prediction. Aging Cell *20*, 1–20.

223. Horvath, S., and Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat. Rev. Genet. *19*, 371–384.

224. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115.

225. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell *49*, 359–367.

226. Jackson, S. H. D., Weale, M. R., and Weale, R. A. (2003). Biological age--what is it and can it be measured? Arch Gerontol Geriatr *36*, 103–115.

227. Levine, M. E., Higgins-Chen, A., Thrush, K., Minteer, C., and Niimi, P. (2022). Clock work: deconstructing the epigenetic clock signals in aging, disease, and reprogramming. BioRxiv.

228. Chen, B. H., Marioni, R. E., Colicino, E., Peters, M. J., Ward-Caviness, C. K., Tsai, P.-C., Roetker, N. S., Just, A. C., Demerath, E. W., Guan, W., et al. (2016). DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany, NY) *8*, 1844–1865.

229. Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., Hou, L., Baccarelli, A. A., Stewart, J. D., Li, Y., et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany, NY) *10*, 573–591.

230. Gross, A. M., Jaeger, P. A., Kreisberg, J. F., Licon, K., Jepsen, K. L., Khosroheidari, M., Morsey, B. M., Swindells, S., Shen, H., Ng, C. T., et al. (2016). Methylome-wide Analysis of Chronic HIV Infection Reveals Five-Year Increase in Biological Age and Epigenetic Targeting of HLA. Mol. Cell *62*, 157–168.

231. Horvath, S., Garagnani, P., Bacalini, M. G., Pirazzini, C., Salvioli, S., Gentilini, D., Di Blasio, A. M., Giuliani, C., Tung, S., Vinters, H. V., et al. (2015). Accelerated epigenetic aging in Down syndrome. Aging Cell *14*, 491–495.

232. Maierhofer, A., Flunkert, J., Oshima, J., Martin, G. M., Haaf, T., and Horvath, S. (2017). Accelerated epigenetic aging in Werner syndrome. Aging (Albany, NY) *9*, 1143–1152.

233. Horvath, S., Pirazzini, C., Bacalini, M. G., Gentilini, D., Di Blasio, A. M., Delledonne, M., Mari, D., Arosio, B., Monti, D., Passarino, G., et al. (2015). Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring. Aging (Albany, NY) *7*, 1159–1170.

234. Levine, M. E., Lu, A. T., Bennett, D. A., and Horvath, S. (2015). Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. Aging (Albany, NY) *7*, 1198–1211.

235. Horvath, S., and Ritz, B. R. (2015). Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. Aging (Albany, NY) *7*, 1130–1142.

236. Loomba, R., Gindin, Y., Jiang, Z., Lawitz, E., Caldwell, S., Djedjos, C. S., Xu, R., Chung, C., Myers, R. P., Subramanian, G. M., et al. (2018). DNA methylation signatures reflect aging in patients with nonalcoholic steatohepatitis. JCI Insight *3*.

237. Marioni, R. E., Shah, S., McRae, A. F., Ritchie, S. J., Muniz-Terrera, G., Harris, S. E., Gibson, J., Redmond, P., Cox, S. R., Pattie, A., et al. (2015). The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. Int. J. Epidemiol. *44*, 1388–1396.

238. Durso, D. F., Bacalini, M. G., Sala, C., Pirazzini, C., Marasco, E., Bonafé, M., do Valle, Í. F., Gentilini, D., Castellani, G., Faria, A. M. C., et al. (2017). Acceleration of leukocytes' epigenetic age as an early tumor and sex-specific marker of breast and colorectal cancer. Oncotarget *8*, 23237–23245.

239. Ambatipudi, S., Horvath, S., Perrier, F., Cuenin, C., Hernandez-Vargas, H., Le Calvez-Kelm, F., Durand, G., Byrnes, G., Ferrari, P., Bouaoun, L., et al. (2017). DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. Eur. J. Cancer *75*, 299–307.

240. Zheng, Y., Joyce, B. T., Colicino, E., Liu, L., Zhang, W., Dai, Q., Shrubsole, M. J., Kibbe, W. A., Gao, T., Zhang, Z., et al. (2016). Blood epigenetic age may predict cancer incidence and mortality. EBioMedicine *5*, 68–73.

241. Bocklandt, S., Lin, W., Sehl, M. E., Sánchez, F. J., Sinsheimer, J. S., Horvath, S., and Vilain, E. (2011). Epigenetic predictor of age. PLoS One *6*, e14821.

242. Koch, C. M., and Wagner, W. (2011). Epigenetic-aging-signature to determine age in different tissues. Aging (Albany, NY) *3*, 1018–1027.

243. Koch, C. M., Joussen, S., Schellenberg, A., Lin, Q., Zenke, M., and Wagner, W. (2012). Monitoring of cellular senescence by DNA-methylation at specific CpG sites. Aging Cell *11*, 366–369.

244. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J Royal Statistical Soc B *67*, 301–320.

245. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological).

246. Levine, M. E. (2013). Modeling the rate of senescence: Can estimated biological age predict mortality more accurately than chronological age? J. Gerontol. A, Biol. Sci. Med. Sci. *68*, 667–674.

247. Lu, A. T., Quach, A., Wilson, J. G., Reiner, A. P., Aviv, A., Raj, K., Hou, L., Baccarelli, A. A., Li, Y., Stewart, J. D., et al. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany, NY) *11*, 303–327.

248. Horvath, S., Gurven, M., Levine, M. E., Trumble, B. C., Kaplan, H., Allayee, H., Ritz, B. R., Chen, B., Lu, A. T., Rickabaugh, T. M., et al. (2016). An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. *17*, 1–22.

249. Quach, A., Levine, M. E., Tanaka, T., Lu, A. T., Chen, B. H., Ferrucci, L., Ritz, B., Bandinelli, S., Neuhouser, M. L., Beasley, J. M., et al. (2017). Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. Aging (Albany, NY) *9*, 419–446.

250. Zhang, Q., Vallerga, C. L., Walker, R. M., Lin, T., Henders, A. K., Montgomery, G. W., He, J., Fan, D., Fowdar, J., Kennedy, M., et al. (2019). Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. Genome Med. *11*, 54.

251. Deary, I. J., Gow, A. J., Taylor, M. D., Corley, J., Brett, C., Wilson, V., Campbell, H., Whalley, L. J., Visscher, P. M., Porteous, D. J., et al. (2007). The Lothian Birth Cohort 1936: a study to examine influences on cognitive ageing from age 11 to age 70 and beyond. BMC Geriatr. *7*, 28.

252. Deary, I. J., Gow, A. J., Pattie, A., and Starr, J. M. (2012). Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. Int. J. Epidemiol. *41*, 1576–1584.

253. 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.

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

255. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

256. Faust, G. G., and Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics *30*, 2503–2505.

257. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

258. Benjamin, D. I., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2. BioRxiv.

259. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*, D941–D947.

260. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome Biol. *17*, 122.

261. Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369.

262. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., and Siegmund, K. D. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. *41*, e90.

263. Chen, Z., Amro, E. M., Becker, F., Hölzer, M., Rasa, S. M. M., Njeru, S. N., Han, B., Di Sanzo, S., Chen, Y., Tang, D., et al. (2019). Cohesin-mediated NF-κB signaling limits hematopoietic stem cell self-renewal in aging and inflammation. J. Exp. Med. *216*, 152–175.

264. Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics *13*, 86.

265. Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data *3*, 160025.

266. 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.

267. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv.

268. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

269. Mahmood, K., Jung, C.-H., Philip, G., Georgeson, P., Chung, J., Pope, B. J., and Park, D. J. (2017). Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum. Genomics *11*, 10.

270. Livesey, B. J., and Marsh, J. A. (2022). Interpreting protein variant effects with computational predictors and deep mutational scanning. Dis. Model. Mech. *15*.

271. 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 Suppl 3*, S3.

272. Hecht, M., Bromberg, Y., and Rost, B. (2015). Better prediction of functional effects for sequence variants. BMC Genomics *16 Suppl 8*, S1.

273. 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.

274. 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.

275. 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.

276. 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.

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

278. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

279. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. *50*, D439–D444.

280. Python 3 Reference Manual: | Guide books Available at: https://dl.acm.org/doi/book/10.5555/1593511 [Accessed May 1, 2023].

281. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.

282. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

283. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. Cell *153*, 1194–1217.

284. Lu, A. T., Xue, L., Salfati, E. L., Chen, B. H., Ferrucci, L., Levy, D., Joehanes, R., Murabito, J. M., Kiel, D. P., Tsai, P.-C., et al. (2018). GWAS of epigenetic aging rates in blood reveals a critical role for TERT. Nat. Commun. *9*, 387.

285. Nachun, D., Lu, A. T., Bick, A. G., Natarajan, P., Weinstock, J., Szeto, M. D., Kathiresan, S., Abecasis, G., Taylor, K. D., Guo, X., et al. (2021). Clonal

hematopoiesis associated with epigenetic aging and clinical outcomes. Aging Cell *20*, e13366.

286. Feldkamp, J. D., Vetter, V. M., Arends, C. M., Lang, T. J. L., Bullinger, L., Damm, F., Demuth, I., and Frick, M. (2022). Clonal hematopoiesis of indeterminate potential-related epigenetic age acceleration correlates with clonal hematopoiesis of indeterminate potential clone size in patients with high morbidity. Haematologica *107*, 1703–1708.

287. Soerensen, M., Tulstrup, M., Hansen, J. W., Weischenfeldt, J., Grønbæk, K., and Christensen, K. (2022). Clonal hematopoiesis and epigenetic age acceleration in elderly danish twins. HemaSphere *6*, e768.

288. de Magalhães, J. P. (2013). How ageing processes influence cancer. Nat. Rev. Cancer *13*, 357–365.

289. Martincorena, I. (2019). Somatic mutation and clonal expansions in human tissues. Genome Med. *11*, 35.

290. 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.

291. 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*, 532–537.

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

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

294. 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.

295. 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*.

296. Abplanalp, W. T., Mas-Peiro, S., Cremer, S., John, D., Dimmeler, S., and Zeiher, A. M. (2020). Association of clonal hematopoiesis of indeterminate potential with inflammatory gene expression in patients with severe degenerative aortic valve stenosis or chronic postischemic heart failure. JAMA Cardiol. *5*, 1170–1175.

297. Ostrander, E. L., Kramer, A. C., Mallaney, C., Celik, H., Koh, W. K., Fairchild, J., Haussler, E., Zhang, C. R. C., and Challen, G. A. (2020). Divergent effects of dnmt3a and tet2 mutations on hematopoietic progenitor cell fitness. Stem Cell Rep. *14*, 551–560.

298. Uddin, M. D. M., Nguyen, N. Q. H., Yu, B., Brody, J. A., Pampana, A., Nakao, T., Fornage, M., Bressler, J., Sotoodehnia, N., Weinstock, J. S., et al. (2022). Clonal

hematopoiesis of indeterminate potential, DNA methylation, and risk for coronary artery disease. Nat. Commun. *13*, 5350.

Appendix 1: List of Unique Variants Detected at 2% VAF

Participant ID	Gene Name	Protein Substitution	Base Substitution	Variant Class	Largest VAF
CHIP_LBC21_011	ASXL1	p.Gln925Ter	c.2773C>T	Nonsense_Mutation	0.0242
CHIP_LBC21_015	ASXL1	p.Val807Ile	c.2419G>A	Missense_Mutation	0.0505
CHIP_LBC21_016	ASXL1	p.Glu635ArgfsTer15	c.1900_1922del	Frame_Shift_Del	0.0572
CHIP_LBC21_017	BCOR	p.Thr1265Pro	c.3793A>C	Missense_Mutation	0.0218
CHIP_LBC21_011	BCORL1	p.Val105Gly	c.314T>G	Missense_Mutation	0.024
CHIP_LBC36_037	BRAF	p.Gly455Arg	c.1363G>A	Missense_Mutation	0.0218
CHIP_LBC21_038	CBL	p.Leu380Pro	c.1139T>C	Missense_Mutation	0.0394
CHIP_LBC36_037	CBL	p.Gly104Arg	c.310G>A	Missense_Mutation	0.023
CHIP_LBC36_037	CBLC	p.Pro438Leu	c.1313C>T	Missense_Mutation	0.0204
CHIP_LBC21_011	CDKN2A	p.Val28Gly	c.83T>G	Missense_Mutation	0.0235
CHIP_LBC21_018	CDKN2A	p.Val82Gly	c.245T>G	Missense_Mutation	0.02
CHIP_LBC36_002	CDKN2A	p.Thr79Pro	c.235A>C	Missense_Mutation	0.0258
CHIP_LBC36_035	CEBPA	p.Tyr181Ser	c.542A>C	Missense_Mutation	0.0204
CHIP_LBC21_002	DNMT3A	p.Ile769Ser	c.2306T>G	Missense_Mutation	0.0221
CHIP_LBC21_003	DNMT3A		c.1668-3C>G	Splice_Region	0.4261
CHIP_LBC21_005	DNMT3A	p.Ile840Ter	c.2517del	Frame_Shift_Del	0.4481
CHIP_LBC21_007	DNMT3A	p.Asp748AlafsTer3	c.2243_2259del	Frame_Shift_Del	0.0931
CHIP_LBC21_007	DNMT3A	p.Asn797Ser	c.2390A>G	Missense_Mutation	0.1423
CHIP_LBC21_013	DNMT3A	p.Ser255ProfsTer61	c.762del	Frame_Shift_Del	0.0377
CHIP_LBC21_015	DNMT3A	p.Val657Met	c.1969G>A	Missense_Mutation	0.0501
CHIP_LBC21_024	DNMT3A	p.Leu504ProfsTer42	c.1510dup	Frame_Shift_Ins	0.1086
CHIP_LBC21_024	DNMT3A	p.Lys744Ter	c.2230A>T	Nonsense_Mutation	0.1097
CHIP_LBC21_032	DNMT3A	p.Val687Asp	c.2060T>A	Missense_Mutation	0.0755
CHIP_LBC21_034	DNMT3A		c.2173+1G>A	Splice_Site	0.0476
CHIP_LBC36_009	DNMT3A	p.Leu547Phe	c.1639C>T	Missense_Mutation	0.297
CHIP_LBC36_011	DNMT3A		c.1852-1G>A	Splice_Site	0.3006
CHIP_LBC36_012	DNMT3A	p.Pro627ArgfsTer22	c.1878_1884del	Frame_Shift_Del	0.1057
CHIP_LBC36_013	DNMT3A	p.Arg635Gln	c.1904G>A	Missense_Mutation	0.16
CHIP_LBC36_014	DNMT3A	p.Gln534ArgfsTer117	c.1601del	Frame_Shift_Del	0.132
CHIP_LBC36_015	DNMT3A	p.Leu344Gln	c.1031T>A	Missense_Mutation	0.0219
CHIP_LBC36_015	DNMT3A	p.Thr257MetfsTer59	c.770del	Frame_Shift_Del	0.1056
CHIP_LBC36_016	DNMT3A	p.Glu725GlyfsTer54	c.2174del	Frame_Shift_Del	0.0596
CHIP_LBC36_017	DNMT3A		c.2322+1G>A	Splice_Site	0.1588
CHIP_LBC36_018	DNMT3A		c.1554+1G>A	Splice_Site	0.0938
CHIP_LBC36_020	DNMT3A	p.Tyr536ThrfsTer115	c.1605del	Frame_Shift_Del	0.2593
CHIP_LBC36_021	DNMT3A	p.Cys559Ter	c.1677C>A	Nonsense_Mutation	0.3837
CHIP_LBC36_021	DNMT3A	p.Arg771Gln	c.2312G>A	Missense_Mutation	0.3827
CHIP_LBC36_022	DNMT3A	p.Arg326Gly	c.976C>G	Missense_Mutation	0.28
CHIP_LBC36_023	DNMT3A		c.1429+1G>C	Splice_Site	0.037
CHIP_LBC36_024	DNMT3A	p.Gly550Arg	c.1648G>A	Missense_Mutation	0.026
CHIP_LBC36_028	DNMT3A		c.2478+1G>T	Splice_Site	0.3834
CHIP_LBC36_030	DNMT3A	p.Phe752LeufsTer4	c.2256_2257del	Frame_Shift_Del	0.3116
CHIP_LBC36_034	DNMT3A	p.Tyr735Asn	c.2203T>A	Missense_Mutation	0.1961
CHIP_LBC36_035	DNMT3A	p.Arg309ProfsTer8	c.924_925dup	Frame_Shift_Ins	0.3821
CHIP_LBC36_036	DNMT3A	p.Gly673ProfsTer35	c.2016_2029del	Frame_Shift_Del	0.2438
CHIP_LBC36_036	DNMT3A	p.Met682IlefsTer31	c.2045dup	Frame_Shift_Ins	0.0415
CHIP_LBC36_037	DNMT3A	p.Met761Ile	c.2283G>A	Missense_Mutation	0.02
CHIP_LBC36_040	DNMT3A		c.2083-2A>G	Splice_Site	0.0432
CHIP_LBC36_041	DNMT3A	p.Arg736Cys	c.2206C>T	Missense_Mutation	0.0936
CHIP_LBC36_042	DNMT3A	p.Ala353ValfsTer39	c.1056_1057del	Frame_Shift_Del	0.3256
CHIP_LBC36_043	DNMT3A	p.Trp581Gly	c.1741T>G	Missense_Mutation	0.0842
CHIP_LBC36_043	DNMT3A	p.Trp860Arg	c.2578T>C	Missense_Mutation	0.0425
CHIP_LBC36_034	EZH2	p.Ser443Ter	c.1328C>A	Nonsense_Mutation	0.024
CHIP_LBC21_017	GATA2	p.Thr457Pro	c.1369A>C	Missense_Mutation	0.0201
CHIP_LBC36_021	GATA2	p.Arg396Trp	c.1186C>T	Missense_Mutation	0.039

Participant ID	Gene Name	Protein Substitution	Base Substitution	Variant Class	Largest VAF
CHIP_LBC21_003	JAK2	p.Val617Phe	c.1849G>T	Missense_Mutation	0.8352
CHIP_LBC21_010	JAK2	p.Val617Phe	c.1849G>T	Missense_Mutation	0.2772
CHIP_LBC36_003	JAK2	p.Val617Phe	c.1849G>T	Missense_Mutation	0.5015
CHIP_LBC36_004	JAK2	p.Val617Phe	c.1849G>T	Missense_Mutation	0.2546
CHIP_LBC36_007	JAK2	p.Val617Phe	c.1849G>T	Missense_Mutation	0.3028
CHIP_LBC36_022	JAK2	p.Val617Phe	c.1849G>T	Missense_Mutation	0.0792
CHIP_LBC36_026	JAK2	p.Val617Phe	c.1849G>T	Missense_Mutation	0.3183
CHIP_LBC36_027	JAK2	p.Val617Phe	c.1849G>T	Missense_Mutation	0.7418
CHIP_LBC36_030	KMT2A	p.Gln33Pro	c.98A>C	Missense_Mutation	0.0226
CHIP_LBC36_031	LUC7L2	p.Met10Val	c.28A>G	Missense_Mutation	0.0248
CHIP_LBC21_017	NF1	p.Ala1197Ser	c.3589G>T	Missense_Mutation	0.1076
CHIP_LBC21_018	NF1	p.Val7Gly	c.20T>G	Missense_Mutation	0.0209
CHIP_LBC36_040	PPM1D	p.Glu459Ter	c.1375G>T	Nonsense_Mutation	0.0204
CHIP_LBC21_004	RAD21	p.Asp400Ala	c.1199A>C	Missense_Mutation	0.0238
CHIP_LBC21_018	RAD21	p.Val140dup	c.419_421dup	In_Frame_Ins	0.0246
CHIP_LBC21_023	RAD21		c.815-3_815-2delinsGG	Splice_Site	0.0206
CHIP_LBC21_037	RUNX1	p.Ser410Ala	c.1228T>G	Missense_Mutation	0.0206
CHIP_LBC36_003	SF3B1	p.Gly742Asp	c.2225G>A	Missense_Mutation	0.0829
CHIP_LBC36_005	SF3B1	p.Lys700Glu	c.2098A>G	Missense_Mutation	0.0257
CHIP_LBC36_037	STAT3	p.Met586Ile	c.1758G>A	Missense_Mutation	0.0215
CHIP_LBC21_001	TET2	p.Asp236IlefsTer14	c.705del	Frame_Shift_Del	0.2107
CHIP_LBC21_008	TET2	p.Glu350LeufsTer24	c.1043_1047dup	Frame_Shift_Ins	0.0319
CHIP_LBC21_008	TET2	p.Ser714Ter	c.2141C>G	Nonsense_Mutation	0.0349
CHIP_LBC21_019	TET2	p.Thr313TyrfsTer18	c.936dup	Frame_Shift_Ins	0.1293
CHIP_LBC21_024	TET2	p.Arg550Ter	c.1648C>T	Nonsense_Mutation	0.1156
CHIP_LBC21_027	TET2	p.Lys780SerfsTer8	c.2339_2340del	Frame_Shift_Del	0.0287
CHIP_LBC21_039	TET2	p.Arg581His	c.1742G>A	Missense_Mutation	0.0308
CHIP_LBC21_046	TET2	p.Pro851LeufsTer22	c.2552del	Frame_Shift_Del	0.132
CHIP_LBC21_046	TET2	p.Gln916Ter	c.2746C>T	Nonsense_Mutation	0.0541
CHIP_LBC36_005	TET2	p.Ser502LeufsTer31	c.1505del	Frame_Shift_Del	0.2661
CHIP_LBC36_008	TET2	p.Glu368AsnfsTer4	c.1102del	Frame_Shift_Del	0.3403
CHIP_LBC36_017	TET2	p.Tyr192His	c.574T>C	Missense_Mutation	0.0667
CHIP_LBC36_019	TET2	p.Tyr192His	c.574T>C	Missense_Mutation	0.4671
CHIP_LBC36_020	TET2	p.Ser153PhefsTer9	c.457dup	Frame_Shift_Ins	0.0391
CHIP_LBC36_024	TET2	p.Asn752ArgfsTer59	c.2255_2261del	Frame_Shift_Del	0.1445
CHIP_LBC36_033	TET2	p.Ile565AsnfsTer2	c.1693dup	Frame_Shift_Ins	0.0368
CHIP_LBC36_033	TET2	p.Ile565Thr	c.1694T>C	Missense_Mutation	0.0388
CHIP_LBC36_043	TET2	p.Phe854LeufsTer19	c.2562del	Frame_Shift_Del	0.082
CHIP_LBC36_003	TP53	p.Glu258Lys	c.772G>A	Missense_Mutation	0.0223
CHIP_LBC21_028	ZRSR2	p.Gly438Arg	c.1311_1312inv	Missense_Mutation	0.0221
CHIP_LBC36_016	ZRSR2	p.Tyr175Cys	c.524A>G	Missense_Mutation	0.0412

Appendix 2: Complete List of Unique Fit CHIP Variants at 2% VAF

Gene	Base Substitution	Protein Substitution	Fitness	Fitness Confidence	Co-Occurring Mutations
RAD21	c.815-3_815-2delinsGG		0.075	[0.01 0.24]	
DNMT3A	c.2173+1G>A		0.02	[0. 0.065]	
DNMT3A	c.1852-1G>A		0.015	[0. 0.05]	
DNMT3A	c.1554+1G>A		0	[0. 0.015]	
DNMT3A	c.1429+1G>C		0	[0. 0.03]	
DNMT3A	c.2083-2A>G		0.055	[0.015 0.1]	PPM1D c.1375G>T
DNMT3A	c.2322+1G>A		0	[0. 0.]	TET2 c.574T>C
DNMT3A	c.2478+1G>T		0.1	[0.025 0.165]	
NF1	c.3589G>T	p.A1197S	0.055	[0.025 0.09]	GATA2 c.1369A>C, BCOR c.3793A>C
DNMT3A	c.1056_1057del	p.A353Vfs*39	0.08	[0.05 0.1]	
DNMT3A	c.1677C>A	p.C559*	0.105	[0.09 0.12]	GATA2 c.1186C>T, DNMT3A c.2312G>A
TET2	c.705del	p.D236Ifs*14	0.155	[0.12 0.19]	
RAD21	c.1199A>C	p.D400A	0	[0. 0.08]	
TET2	c.1102del	p.E368Nfs*4	0.185	[0.17 0.2]	
PPM1D	c.1375G>T	p.E459*	0.055	[0.015 0.1]	DNMT3A c.2083-2A>G
ASXL1	c.1900_1922del	p.E635Rfs*15	0.135	[0.1 0.175]	
DNMT3A	c.2174del	p.E725Gfs*54	0.04	[0.005 0.1]	ZRSR2 c.524A>G
DNMT3A	c.2256_2257del	p.F752Lfs*4	0.09	[0.06 0.125]	KMT2A c.98A>C
TET2	c.2562del	p.F854Lfs*19	0	[0. 0.]	DNMT3A c.1741T>G, DNMT3A c.2578T>C
ZRSR2	c.1311_1312inv	p.G438R	0	[0. 0.055]	
DNMT3A	c.1648G>A	p.G550R	0.105	[0.08 0.13]	TET2 c.2255_2261del
DNMT3A	c.2016_2029del	p.G673Pfs*35	0	[0. 0.005]	DNMT3A c.2045dup
TET2	c.1693dup	p.I565Nfs*2	0.05	[0.03 0.075]	TET2 c.1694T>C
TET2	c.1694T>C	p.I565T	0.05	[0.03 0.075]	TET2 c.1693dup
DNMT3A	c.2306T>G	p.I769S	0	[0. 0.07]	
SF3B1	c.2098A>G	p.K700E	0.485	[0.375 0.595]	TET2 c.1505del
DNMT3A	c.2230A>T	p.K744*	0.06	[0.025 0.09]	DNMT3A c.1510dup, TET2 c.1648C>T
TET2	c.2339_2340del	p.K780Sfs*8	0	[0. 0.05]	
DNMT3A	c.1031T>A	p.L344Q	0	[0. 0.025]	DNMT3A c.770del
CBL	c.1139T>C	p.L380P	0	[0. 0.035]	
DNMT3A	c.1510dup	p.L504Pfs*42	0.06	[0.025 0.09]	TET2 c.1648C>T, DNMT3A c.2230A>T
DNMT3A	c.1639C>T	p.L547F	0.025	[0.01 0.045]	
LUC7L2	c.28A>G	p.M10V	0.205	[0.035 0.52]	
DNMT3A	c.2045dup	p.M682Ifs*31	0	[0. 0.005]	DNMT3A c.2016_2029del
TET2	c.2255_2261del	p.N752Rfs*59	0.105	[0.08 0.13]	DNMT3A c.1648G>A
DNMT3A	c.1878_1884del	p.P627Rfs*22	0	[0. 0.01]	
TET2	c.2552del	p.P851Lfs*22	0.135	[0.095 0.175]	TET2 c.2746C>T
KMT2A	c.98A>C	p.Q33P	0.09	[0.06 0.125]	DNMT3A c.2256_2257del
DNMT3A	c.1601del	p.Q534Rfs*117	0.02	[0.005 0.035]	
TET2	c.2746C>T	p.Q916*	0	[0. 0.06]	TET2 c.2552del
ASXL1	c.2773C>T	p.Q925*	0.185	[0.12 0.255]	CDKN2A c.83T>G, BCORL1 c.314T>G
DNMT3A	c.924_925dup	p.R309Pfs*8	0.095	[0.055 0.14]	CEBPA c.542A>C
DNMT3A	c.976C>G	p.R326G	0.14	[0.115 0.165]	JAK2 c.1849G>T
GATA2	c.1186C>T	p.R396W	0.31	[0.27 0.335]	DNMT3A c.1677C>A, DNMT3A c.2312G>A
TET2	c.1648C>T	p.R550*	0.06	[0.025 0.09]	DNMT3A c.1510dup, DNMT3A c.2230A>T
TET2	c.1742G>A	p.R581H	0.15	[0.1 0.205]	

Gene	Base Substitution	Protein Substitution	Fitness	Fitness Confidence	Co-Occurring Mutations
DNMT3A	c.1904G>A	p.R635Q	0.175	[0.14 0.205]	
DNMT3A	c.2206C>T	p.R736C	0.05	[0.03 0.075]	
DNMT3A	c.2312G>A	p.R771Q	0.105	[0.09 0.12]	GATA2 c.1186C>T, DNMT3A c.1677C>A
TET2	c.457dup	p.S153Ffs*9	0.1	[0.08 0.12]	DNMT3A c.1605del
RUNX1	c.1228T>G	p.S410A	0	[0. 0.095]	
EZH2	c.1328C>A	p.S443*	0.235	[0.16 0.33]	DNMT3A c.2203T>A
TET2	c.1505del	p.S502Lfs*31	0.145	[0.125 0.165]	SF3B1 c.2098A>G
BCOR	c.3793A>C	p.T1265P	0.055	[0.025 0.09]	NF1 c.3589G>T, GATA2 c.1369A>C
DNMT3A	c.770del	p.T257Mfs*59	0	[0. 0.025]	DNMT3A c.1031T>A
TET2	c.936dup	p.T313Yfs*18	0.335	[0.28 0.39]	
GATA2	c.1369A>C	p.T457P	0.055	[0.025 0.09]	NF1 c.3589G>T, BCOR c.3793A>C
CDKN2A	c.235A>C	p.T79P	0.065	[0.02 0.125]	
BCORL1	c.314T>G	p.V105G	0	[0. 0.04]	CDKN2A c.83T>G, ASXL1 c.2773C>T
RAD21	c.419_421dup	p.V140dup	0.08	[0.03 0.135]	NF1 c.20T>G
CDKN2A	c.83T>G	p.V28G	0	[0. 0.04]	BCORL1 c.314T>G, ASXL1 c.2773C>T
JAK2	c.1849G>T	p.V617F	0.145	[0.13 0.16]	
JAK2	c.1849G>T	p.V617F	0.015	[0. 0.075]	
JAK2	c.1849G>T	p.V617F	0.14	[0.115 0.165]	DNMT3A c.976C>G
JAK2	c.1849G>T	p.V617F	0.12	[0.105 0.14]	
DNMT3A	c.1969G>A	p.V657M	0.01	[0. 0.08]	ASXL1 c.2419G>A
DNMT3A	c.2060T>A	p.V687D	0.065	[0.02 0.115]	
NF1	c.20T>G	p.V7G	0.08	[0.03 0.135]	RAD21 c.419_421dup
ASXL1	c.2419G>A	p.V807I	0.01	[0. 0.08]	DNMT3A c.1969G>A
DNMT3A	c.1741T>G	p.W581G	0	[0. 0.]	DNMT3A c.2578T>C, TET2 c.2562del
DNMT3A	c.2578T>C	p.W860R	0.115	[0.02 0.245]	DNMT3A c.1741T>G, TET2 c.2562del
ZRSR2	c.524A>G	p.Y175C	0.04	[0.005 0.1]	DNMT3A c.2174del
CEBPA	c.542A>C	p.Y181S	0.095	[0.055 0.14]	DNMT3A c.924_925dup
TET2	c.574T>C	p.Y192H	0	[0. 0.]	DNMT3A c.2322+1G>A
DNMT3A	c.1605del	p.Y536Tfs*115	0.175	[0.155 0.195]	TET2 c.457dup
DNMT3A	c.2203T>A	p.Y735N	0.105	[0.09 0.125]	EZH2 c.1328C>A

Appendix 3: LiFT-Filter Variant Fitness Estimates

Participant ID	Gene	Base Substitution	Protein ID	Exceeds 2% VAF	Fitness	Fitness Confidence	Co-Occuring Mutation
LBC21_018	BCORL1	c.4619-5A>G		FALSE	0.07	[0.03 0.115]	NF1 c.20T>G, RAD21 c.419_421dup, BCORL1 c.707T>G
LBC21_016	ETV6	c.34-4A>G		FALSE	0.525	[0.25 0.77]	ASXL1 c.1900_1922del
LBC21_030	BCOR	c.3503-51A>C		FALSE	0.45	[0.32 0.6]	TET2 c.961C>T
LBC21_034	STAG2	c.463-1G>A		FALSE	0.03	[0.005 0.07]	DNMT3A c.2173+1G>A
LBC21_034	DNMT3A	c.2173+1G>A		TRUE	0.03	[0.005 0.07]	STAG2 c.463-1G>A
LBC36_020	TP53	c.994-5T>C		FALSE	0.265	[0.215 0.32]	LUC7L2 c.279del, DNMT3A c.1605del, TET2 c.457dup
LBC36_017	DNMT3A	c.2322+1G>A		TRUE	0.06	[0.04 0.08]	
LBC36_018	DNMT3A	c.1554+1G>A		TRUE	0	[0. 0.015]	
LBC36_011	DNMT3A	c.1852-1G>A		TRUE	0.015	[0. 0.05]	TP53 c.818G>A
LBC36_038	KDM6A	c.1329+457G>T		FALSE	0.145	[0.1 0.19]	LUC7L2 c.157A>G
LBC36_040	DNMT3A	c.2083-2A>G		TRUE	0.06	[0.02 0.105]	PPM1D c.1375G>T, SRSF2 c.330C>G
LBC36_023	DNMT3A	c.1429+1G>C		TRUE	0	[0. 0.03]	KRAS c.436G>C, TP53 c.711G>A
LBC36_028	DNMT3A	c.2478+1G>T		TRUE	0.105	[0.04 0.18]	TP53 c.757_758dup
LBC36_043	DNMT3A	c.1122+1G>A		FALSE	0.47	[0.215 0.79]	DNMT3A c.2727T>A, DNMT3A c.2578T>C, DNMT3A c.2228C>T
LBC21_040	SH2B3	c.331G>C	p.A111P	FALSE	0.27	[0.135 0.43]	
LBC21_017	NF1	c.3589G>T	p.A1197S	TRUE	0.075	[0.04 0.115]	
LBC36_023	KRAS	c.436G>C	p.A146P	FALSE	0.175	[0.105 0.24]	DNMT3A c.1429+1G>C, TP53 c.711G>A
LBC21_038	KDM6A	c.51T>G	p.A17=	FALSE	0.515	[0.26 0.635]	CBL c.1139T>C, SRSF2 c.361A>G
LBC36_042	DNMT3A	c.1056_1057del	p.A353Vfs*39	TRUE	0.08	[0.055 0.1]	BCORL1 c.1261T>C
LBC36_042	BCORL1	c.1261T>C	p.C421R	FALSE	0.08	[0.055 0.1]	DNMT3A c.1056_1057del
LBC36_021	DNMT3A	c.1677C>A	p.C559*	TRUE	0.105	[0.09 0.12]	U2AF2 c.1421T>C, GATA2 c.1186C>T, DNMT3A c.2312G>A
LBC21_001	TET2	c.705del	p.D236Ifs*14	TRUE	0.155	[0.12 0.19]	
LBC21_004	RAD21	c.1199A>C	p.D400A	TRUE	0	[0. 0.085]	BCOR c.1163T>G, ATRX c.6524T>C
LBC36_008	TET2	c.1102del	p.E368Nfs*4	TRUE	0.185	[0.17 0.2]	
LBC36_040	PPM1D	c.1375G>T	p.E459*	TRUE	0.06	[0.02 0.105]	SRSF2 c.330C>G, DNMT3A c.2083-2A>G
LBC21_016	ASXL1	c.1900_1922del	p.E635Rfs*15	TRUE	0.135	[0.105 0.18]	ETV6 c.34-4A>G
LBC36_016	DNMT3A	c.2174del	p.E725Gfs*54	TRUE	0.04	[0.005 0.1]	ZRSR2 c.524A>G
LBC36_006	ASXL1	c.2329G>T	p.E777*	FALSE	0.06	[0.02 0.105]	DNMT3A c.1156del
LBC21_011	PTEN	c.834C>G	p.F278L	FALSE	0.16	[0.115 0.21]	ASXL1 c.2773C>T
LBC36_021	U2AF2	c.1421T>C	p.F474S	FALSE	0.31	[0.275 0.34]	GATA2 c.1186C>T, DNMT3A c.1677C>A, DNMT3A c.2312G>A
LBC21_021	PPM1D	c.1602del	p.F534Lfs*5	FALSE	0.13	[0.04 0.23]	
LBC36_030	DNMT3A	c.2256_2257del	p.F752Lfs*4	TRUE	0.095	[0.065 0.13]	NF1 c.2935A>C

Participant ID	Gene	Base Substitution	Protein ID	Exceeds 2% VAF	Fitness	Fitness Confidence	Co-Occuring Mutation
LBC36_043	DNMT3A	c.2727T>A	p.F909L	FALSE	0.195	[0.105 0.29]	DNMT3A c.1122+1G>A, DNMT3A c.2578T>C, DNMT3A c.2228C>T
LBC36_020	LUC7L2	c.279del	p.F93Lfs*16	FALSE	0.265	[0.215 0.32]	DNMT3A c.1605del, TET2 c.457dup, TP53 c.994-5T>C
LBC21_028	ZRSR2	c.1311_1312inv	p.G438R	TRUE	0.005	[0. 0.055]	STAG2 c.2725C>T
LBC36_024	DNMT3A	c.1648G>A	p.G550R	TRUE	0.185	[0.135 0.245]	DNMT3A c.2330C>G, TET2 c.2255_2261del
LBC36_036	DNMT3A	c.2016_2029del	p.G673Pfs*35	TRUE	0	[0. 0.005]	DNMT3A c.2045dup
LBC21_004	ATRX	c.6524T>C	p.I2175T	FALSE	0.28	[0.14 0.43]	RAD21 c.1199A>C, BCOR c.1163T>G
LBC36_028	TP53	c.757_758dup	p.I254Pfs*92	FALSE	0.235	[0.165 0.31]	DNMT3A c.2478+1G>T
LBC36_031	DNMT3A	c.929T>C	p.I310T	FALSE	0.115	[0.035 0.225]	LUC7L2 c.28A>G, CUX1 c.599A>G
LBC36_033	TET2	c.1693dup	p.I565Nfs*2	TRUE	0.055	[0.03 0.08]	U2AF2 c.959T>C, TET2 c.1694T>C
LBC36_033	TET2	c.1694T>C	p.I565T	TRUE	0.055	[0.03 0.08]	U2AF2 c.959T>C, TET2 c.1693dup
LBC21_002	DNMT3A	c.2306T>G	p.I769S	TRUE	0	[0. 0.07]	
LBC21_027	SF3B1	c.2429T>C	p.I810T	FALSE	0.345	[0.11 0.67]	TET2 c.2339_2340del
LBC36_031	CUX1	c.599A>G	p.K200R	FALSE	0.37	[0.175 0.63]	LUC7L2 c.28A>G, DNMT3A c.929T>C
LBC36_005	SF3B1	c.2098A>G	p.K700E	TRUE	0.485	[0.375 0.6]	TET2 c.1505del
LBC21_024	DNMT3A	c.2230A>T	p.K744*	TRUE	0.06	[0.025 0.09]	DNMT3A c.1510dup, TET2 c.1648C>T
LBC21_027	TET2	c.2339_2340del	p.K780Sfs*8	TRUE	0	[0. 0.05]	SF3B1 c.2429T>C
LBC21_036	SH2B3	c.371T>A	p.L124Q	FALSE	0.06	[0.01 0.245]	SH2B3 c.380C>G
LBC36_002	NOTCH1	c.6881T>G	p.L2294R	FALSE	0.235	[0.08 0.425]	
LBC21_018	BCORL1	c.707T>G	p.L236R	FALSE	0.365	[0.115 0.68]	BCORL1 c.4619-5A>G, NF1 c.20T>G, RAD21 c.419_421dup
LBC36_033	U2AF2	c.959T>C	p.L320P	FALSE	0.26	[0.125 0.4]	TET2 c.1693dup, TET2 c.1694T>C
LBC36_015	DNMT3A	c.1031T>A	p.L344Q	TRUE	0	[0. 0.025]	DNMT3A c.770del
LBC21_038	CBL	c.1139T>C	p.L380P	TRUE	0	[0. 0.035]	SRSF2 c.361A>G, KDM6A c.51T>G
LBC21_004	BCOR	c.1163T>G	p.L388R	FALSE	0.28	[0.14 0.43]	RAD21 c.1199A>C, ATRX c.6524T>C
LBC21_024	DNMT3A	c.1510dup	p.L504Pfs*42	TRUE	0.06	[0.025 0.09]	TET2 c.1648C>T, DNMT3A c.2230A>T
LBC36_009	DNMT3A	c.1639C>T	p.L547F	TRUE	0.025	[0.01 0.045]	NF1 c.20T>G
LBC21_006	TET2	c.2243del	p.L748Yfs*3	FALSE	0.08	[0.015 0.165]	
LBC36_031	LUC7L2	c.28A>G	p.M10V	TRUE	0.115	[0.035 0.225]	DNMT3A c.929T>C, CUX1 c.599A>G
LBC36_023	TP53	c.711G>A	p.M237I	FALSE	0.175	[0.105 0.24]	DNMT3A c.1429+1G>C, KRAS c.436G>C
LBC21_045	RAD21	c.860T>C	p.M287T	FALSE	0.005	[0. 0.1]	
LBC36_036	DNMT3A	c.2045dup	p.M682Ifs*31	TRUE	0	[0. 0.005]	DNMT3A c.2016_2029del
LBC36_024	TET2	c.2255_2261del	p.N752Rfs*59	TRUE	0.105	[0.08 0.13]	DNMT3A c.2330C>G, DNMT3A c.1648G>A
LBC21_036	SH2B3	c.380C>G	p.P127R	FALSE	0.06	[0.01 0.245]	SH2B3 c.371T>A
LBC36_012	DNMT3A	c.1878_1884del	p.P627Rfs*22	TRUE	0	[0. 0.01]	
LBC36_043	DNMT3A	c.2228C>T	p.P743L	FALSE	0.195	[0.105 0.29]	DNMT3A c.2727T>A, DNMT3A c.1122+1G>A, DNMT3A c.2578T>C

Participant ID	Gene	Base Substitution	Protein ID	Exceeds 2% VAF	Fitness	Fitness Confidence	Co-Occuring Mutation
LBC36_024	DNMT3A	c.2330C>G	p.P777R	FALSE	0.185	[0.135 0.245]	TET2 c.2255_2261del, DNMT3A c.1648G>A
LBC21_046	TET2	c.2552del	p.P851Lfs*22	TRUE	0.135	[0.095 0.175]	TET2 c.2746C>T
LBC21_019	U2AF1	c.470A>G	p.Q157R	FALSE	0.335	[0.285 0.39]	TET2 c.936dup
LBC21_030	TET2	c.961C>T	p.Q321*	FALSE	0.45	[0.32 0.6]	BCOR c.3503-51A>C
LBC36_014	DNMT3A	c.1601del	p.Q534Rfs*117	TRUE	0.02	[0.005 0.035]	
LBC21_028	STAG2	c.2725C>T	p.Q909*	FALSE	0.145	[0.07 0.23]	ZRSR2 c.1311_1312inv
LBC21_046	TET2	c.2746C>T	p.Q916*	TRUE	0	[0. 0.06]	TET2 c.2552del
LBC21_011	ASXL1	c.2773C>T	p.Q925*	TRUE	0.16	[0.115 0.21]	PTEN c.834C>G
LBC36_041	TP53	c.818G>A	p.R273H	FALSE	0.255	[0.2 0.315]	DNMT3A c.34A>C, DNMT3A c.2206C>T
LBC36_011	TP53	c.818G>A	p.R273H	FALSE	0.015	[0. 0.05]	DNMT3A c.1852-1G>A
LBC36_035	DNMT3A	c.924_925dup	p.R309Pfs*8	TRUE	0.08	[0.045 0.11]	
LBC36_022	DNMT3A	c.976C>G	p.R326G	TRUE	0.15	[0.13 0.175]	JAK2 c.1849G>T, DNMT3A c.2204A>G
LBC36_021	GATA2	c.1186C>T	p.R396W	TRUE	0.31	[0.275 0.34]	U2AF2 c.1421T>C, DNMT3A c.1677C>A, DNMT3A c.2312G>A
LBC36_038	LUC7L2	c.157A>G	p.R53G	FALSE	0.145	[0.1 0.19]	KDM6A c.1329+457G>T
LBC21_024	TET2	c.1648C>T	p.R550*	TRUE	0.06	[0.025 0.09]	DNMT3A c.1510dup, DNMT3A c.2230A>T
LBC21_039	TET2	c.1742G>A	p.R581H	TRUE	0.15	[0.105 0.205]	
LBC36_013	DNMT3A	c.1904G>A	p.R635Q	TRUE	0.175	[0.145 0.205]	JAK2 c.1849G>T
LBC36_041	DNMT3A	c.2206C>T	p.R736C	TRUE	0.06	[0.035 0.08]	DNMT3A c.34A>C, TP53 c.818G>A
LBC36_021	DNMT3A	c.2312G>A	p.R771Q	TRUE	0.105	[0.09 0.12]	U2AF2 c.1421T>C, GATA2 c.1186C>T, DNMT3A c.1677C>A
LBC36_001	DNMT3A	c.2645G>A	p.R882H	FALSE	0.16	[0.08 0.245]	
LBC21_038	SRSF2	c.361A>G	p.S121G	FALSE	0.515	[0.26 0.635]	CBL c.1139T>C, KDM6A c.51T>G
LBC36_020	TET2	c.457dup	p.S153Ffs*9	TRUE	0.11	[0.09 0.13]	LUC7L2 c.279del, DNMT3A c.1605del, TP53 c.994-5T>C
LBC36_034	EZH2	c.1328C>A	p.S443*	TRUE	0.235	[0.16 0.33]	DNMT3A c.34A>C, DNMT3A c.2203T>A
LBC36_005	TET2	c.1505del	p.S502Lfs*31	TRUE	0.145	[0.125 0.165]	SF3B1 c.2098A>G
LBC36_010	FBXW7	c.1547C>T	p.S516F	FALSE	0	[0. 0.12]	
LBC36_030	NF1	c.2935A>C	p.S979R	FALSE	0.43	[0.26 0.655]	DNMT3A c.2256_2257del
LBC36_034	DNMT3A	c.34A>C	p.T12P	FALSE	0.105	[0.085 0.12]	EZH2 c.1328C>A, DNMT3A c.2203T>A
LBC36_041	DNMT3A	c.34A>C	p.T12P	FALSE	0.255	[0.2 0.315]	TP53 c.818G>A, DNMT3A c.2206C>T
LBC36_015	DNMT3A	c.770del	p.T257Mfs*59	TRUE	0	[0. 0.025]	DNMT3A c.1031T>A
LBC21_019	TET2	c.936dup	p.T313Yfs*18	TRUE	0.335	[0.285 0.39]	U2AF1 c.470A>G
LBC21_018	RAD21	c.419_421dup	p.V140dup	TRUE	0.07	[0.03 0.115]	BCORL1 c.4619-5A>G, NF1 c.20T>G, BCORL1 c.707T>G
LBC36_006	DNMT3A	c.1156del	p.V386Cfs*21	FALSE	0.06	[0.02 0.105]	ASXL1 c.2329G>T
LBC36_022	JAK2	c.1849G>T	p.V617F	TRUE	0.15	[0.13 0.175]	DNMT3A c.976C>G, DNMT3A c.2204A>G
LBC21_010	JAK2	c.1849G>T	p.V617F	TRUE	0.015	[0. 0.075]	
LBC36_013	JAK2	c.1849G>T	p.V617F	FALSE	0.075	[0.015 0.155]	DNMT3A c.1904G>A
LBC36_026	JAK2	c.1849G>T	p.V617F	TRUE	0.145	[0.135 0.16]	

Participant ID	Gene	Base Substitution	Protein ID	Exceeds 2% VAF	Fitness	Fitness Confidence	Co-Occuring Mutation
LBC36_007	JAK2	c.1849G>T	p.V617F	TRUE	0.12	[0.105 0.14]	
LBC21_015	DNMT3A	c.1969G>A	p.V657M	TRUE	0.01	[0. 0.08]	ASXL1 c.2419G>A
LBC21_032	DNMT3A	c.2060T>A	p.V687D	TRUE	0.065	[0.02 0.115]	
LBC21_018	NF1	c.20T>G	p.V7G	TRUE	0.07	[0.03 0.115]	BCORL1 c.4619-5A>G, RAD21 c.419_421dup, BCORL1 c.707T>G
LBC36_009	NF1	c.20T>G	p.V7G	FALSE	0.025	[0.01 0.045]	DNMT3A c.1639C>T
LBC21_015	ASXL1	c.2419G>A	p.V807I	TRUE	0.01	[0. 0.08]	DNMT3A c.1969G>A
LBC36_043	DNMT3A	c.2578T>C	p.W860R	TRUE	0.195	[0.105 0.29]	DNMT3A c.2727T>A, DNMT3A c.1122+1G>A, DNMT3A c.2228C>T
LBC36_040	SRSF2	c.330C>G	p.Y110*	FALSE	0.565	[0.225 0.935]	PPM1D c.1375G>T, DNMT3A c.2083-2A>G
LBC36_016	ZRSR2	c.524A>G	p.Y175C	TRUE	0.04	[0.005 0.1]	DNMT3A c.2174del
LBC36_020	DNMT3A	c.1605del	p.Y536Tfs*115	TRUE	0.185	[0.17 0.205]	LUC7L2 c.279del, TET2 c.457dup, TP53 c.994-5T>C
LBC36_022	DNMT3A	c.2204A>G	p.Y735C	FALSE	0.35	[0.255 0.45]	DNMT3A c.976C>G, JAK2 c.1849G>T
LBC36_034	DNMT3A	c.2203T>A	p.Y735N	TRUE	0.105	[0.085 0.12]	EZH2 c.1328C>A, DNMT3A c.34A>C

Appendix 4: Damage Predictions for Single Nucleotide Variants

Gene	Uniprot	Variant	DeepSeq	REVEL	>gnomAD	>ClinVar	SIFT4G	SNAP2	DEOGEN2	VEST4	gnomAD mean	clinVar mean	ΔΔG full	ΔΔG AlphaFold	Category Assigned
GATA2	P23769	R396W	-8.634	0.912	0.967	0.171	0.000	78.000	0.929	0.919	0.970	0.325	-0.116	0.302	
U2AF1	Q01081	Q157R	-	0.596	0.917	0.250	0.001	-	-	0.825	0.907	0.250	-	0.836	possibly damaging
SF3B1	O75533	K700E	-16.139	0.619	0.906	0.400	0.011	53.000	0.587	0.688	0.865	0.433	-0.920	-0.748	
LUC7L2	Q9Y383	R53G	-	0.358	0.960	-	0.002	86.000	0.488	0.816	0.960	-	-	1.269	likely damaging
LUC7L2	Q9Y383	M10V	-	0.202	0.784	-	0.042	46.000	0.212	0.462	0.770	-	-	1.410	
BCOR	Q6W2J9	L388R	-	0.108	0.372	0.000	0.636	-9.000	0.263	0.473	0.452	0.100	-	0.617	likely benign
JAK2	O60674	V617F	-8.876	0.881	0.960	0.857	0.007	88.000	0.845	0.885	0.909	0.589	1.828	1.608	likely damaging
ATRX	P46100	I2175T	-	0.963	1.000	0.826	0.001	-	-	0.866	0.991	0.630	-	1.625	likely damaging
TET2	Q6N021	I565T	-	0.226	0.859	0.500	0.133	-52.000	0.084	0.071	0.456	0.200	-	1.225	
TET2	Q6N021	R581H	-	0.032	0.143	0.000	0.140	-7.000	0.022	0.228	0.363	0.300	-	0.711	
DNMT3A	Q9Y6K1	L344Q	-6.753	0.928	0.882	0.703	0.001	74.000	0.885	0.976	0.829	0.685	2.986	3.536	
DNMT3A	Q9Y6K1	L547F	-6.757	0.653	0.555	0.162	0.139	28.000	0.532	0.763	0.544	0.180	4.604	4.462	
DNMT3A	Q9Y6K1	G550R	-7.397	0.799	0.670	0.324	0.019	46.000	0.816	0.765	0.657	0.338	1.869	-0.265	
DNMT3A	Q9Y6K1	R635Q	-7.159	0.828	0.698	0.351	0.065	39.000	0.946	0.778	0.679	0.383	0.857	1.217	
DNMT3A	Q9Y6K1	V657M	-4.779	0.697	0.585	0.162	0.009	39.000	0.955	0.920	0.701	0.414	2.018	0.945	
DNMT3A	Q9Y6K1	V687D	-5.736	0.952	0.934	0.838	0.000	83.000	0.987	0.958	0.883	0.748	3.925	4.793	
DNMT3A	Q9Y6K1	Y735N	-9.957	0.860	0.742	0.405	0.003	50.000	0.938	0.972	0.826	0.635	2.228	0.720	
DNMT3A	Q9Y6K1	Y735C	-7.865	0.895	0.799	0.514	0.002	23.000	0.939	0.953	0.789	0.577	1.804	0.776	likely damaging
DNMT3A	Q9Y6K1	R736C	-6.080	0.923	0.860	0.676	0.015	36.000	0.987	0.822	0.730	0.477	4.011	2.946	
DNMT3A	Q9Y6K1	P743L	-9.984	0.885	0.786	0.459	0.001	52.000	0.982	0.910	0.841	0.653	2.897	2.276	likely damaging
DNMT3A	Q9Y6K1	I769S	-7.962	0.964	0.962	0.892	0.001	64.000	0.928	0.962	0.867	0.730	3.823	4.410	
DNMT3A	Q9Y6K1	R771Q	-6.601	0.904	0.808	0.541	0.034	-7.000	0.822	0.916	0.643	0.347	-0.061	0.309	
DNMT3A	Q9Y6K1	P777R	-7.739	0.937	0.898	0.757	0.001	47.000	0.872	0.950	0.804	0.640	7.769	7.363	likely damaging
DNMT3A	Q9Y6K1	W860R	-5.792	0.905	0.813	0.595	0.015	87.000	0.842	0.907	0.746	0.536	4.660	2.229	
DNMT3A	Q9Y6K1	R882H	-8.980	0.742	0.621	0.243	0.050	43.000	0.818	0.648	0.633	0.320	-0.194	0.305	possibly damaging
DNMT3A	Q9Y6K1	F909L	-6.782	0.883	0.783	0.432	0.011	63.000	0.806	0.772	0.690	0.356	4.014	2.597	possibly damaging
DNMT3A	Q9Y6K1	I310T	-5.851	0.854	0.731	0.405	0.002	61.000	0.852	0.945	0.744	0.477	1.642	3.642	possibly damaging
DNMT3A	Q9Y6K1	R326G	-5.945	0.759	0.632	0.270	0.004	80.000	0.805	0.973	0.740	0.482	5.912	6.570	
BCORL1	Q5H9F3	C421R	-	-	-	-	-	44.000	-	-	0.944	1.000	-	-0.424	likely damaging
NF1	P21359	S979R	-	0.239	0.352	0.046	0.533	-86.000	0.613	0.710	0.295	0.042	-	-	likely benign
NF1	P21359	A1197S	-	0.291	0.472	0.064	0.022	-74.000	0.667	0.693	0.448	0.071	-	-	
PTPN11	Q06124	K260R	-0.615	0.397	0.332	0.009	0.442	-82.000	0.393	0.647	0.245	0.031	0.058	0.022	likely benign

Gene	Uniprot	Variant	DeepSeq	REVEL	>gnomAD	>ClinVar	SIFT4G	SNAP2	DEOGEN2	VEST4	gnomAD mean	clinVar mean	ΔΔG full	ΔΔG AlphaFold	Category Assigned
PTEN	P60484	F278L	-7.509	0.804	0.831	0.099	0.008	49.000	0.939	0.922	0.898	0.226	1.597	1.193	possibly damaging
KRAS	P01116	A146P	-13.991	0.935	1.000	0.907	0.001	75.000	0.938	0.970	0.971	0.764	-0.948	-1.298	likely damaging
TP53	P04637	M237I	-6.257	0.923	0.906	0.592	0.023	45.000	0.995	0.937	0.853	0.441	2.735	1.230	possibly damaging
TP53	P04637	R273H	-4.405	0.868	0.816	0.218	0.032	72.000	0.995	0.963	0.838	0.395	1.097	-0.315	possibly damaging
RAD21	O60216	D400A	-	0.153	0.474	0.000	0.144	-	0.165	0.305	0.536	0.036	-	0.256	
RAD21	O60216	M287T	-	0.094	0.275	0.000	0.545	-	0.056	0.303	0.229	0.000	-	-0.263	likely benign
ASXL1	Q8IXJ9	V807I	-3.846	0.026	0.180	0.000	0.858	-84.000	0.058	0.026	0.149	0.167	-	0.254	
CBL	P22681	L380P	-	0.974	0.998	0.867	0.001	-	0.965	0.979	0.967	0.583	3.298	0.271	
U2AF2	P26368	F474S	-14.096	0.240	0.890	-	0.001	54.000	0.349	0.839	0.960	-	4.395	4.461	likely damaging
U2AF2	P26368	E71G	-0.627	0.072	0.143	-	0.355	-27.000	0.047	0.471	0.370	-	-	0.474	possibly damaging
U2AF2	P26368	L320P	-14.907	0.841	1.000	-	0.001	86.000	0.362	0.922	0.993	-	8.166	7.678	likely damaging
ZRSR2	Q15696	G438R	-	0.209	0.655	-	0.751	-46.000	0.007	0.102	0.197	-	-	1.395	
ZRSR2	Q15696	Y175C	-	0.590	0.915	-	0.001	73.000	0.451	0.760	0.955	-	-	3.349	
SH2B3	Q9UQQ2	A111P	1.162	0.065	0.311	0.000	0.193	-74.000	0.197	0.082	0.142	0.000	-	2.128	likely benign
SH2B3	Q9UQQ2	L124Q	-2.326	0.187	0.651	0.000	0.017	-38.000	0.483	0.175	0.482	0.083	-	-0.801	likely benign
SH2B3	Q9UQQ2	P127R	-1.527	0.039	0.151	0.000	0.133	-70.000	0.233	0.105	0.200	0.000	-	-2.160	likely benign
SRSF2	Q01130	S121G	-0.531	0.127	0.483	0.000	0.166	11.000	0.070	0.308	0.297	0.500	-	-0.146	possibly damaging

Appendix 5: Manuscript – Longitudinal Dynamics of Clonal Haematopoiesis Identifies Gene-Specific Fitness Effects (2022)

nature medicine

ARTICLES

https://doi.org/10.1038/s41591-022-01883-3

Check for updates

OPEN

Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects

Neil A. Robertson[1,9], Eric Latorre-Crespo[1,9], Maria Terradas-Terradas[2,3], Jorge Lemos-Portela[4], Alison C. Purcell[2,3], Benjamin J. Livesey[1], Robert F. Hillary[5], Lee Murphy[6], Angie Fawkes[6], Louise MacGillivray[6], Mhairi Copland[7], Riccardo E. Marioni[5], Joseph A. Marsh[1], Sarah E. Harris[8], Simon R. Cox[8], Ian J. Deary[8], Linus J. Schumacher[4,9] ✉, Kristina Kirschner[2,3,9] ✉ and Tamir Chandra[1,9] ✉

Clonal hematopoiesis of indeterminate potential (CHIP) increases rapidly in prevalence beyond age 60 and has been associated with increased risk for malignancy, heart disease and ischemic stroke. CHIP is driven by somatic mutations in hematopoietic stem and progenitor cells (HSPCs). Because mutations in HSPCs often drive leukemia, we hypothesized that HSPC fitness substantially contributes to transformation from CHIP to leukemia. HSPC fitness is defined as the proliferative advantage over cells carrying no or only neutral mutations. If mutations in different genes lead to distinct fitness advantages, this could enable patient stratification. We quantified the fitness effects of mutations over 12 years in older age using longitudinal sequencing and developed a filtering method that considers individual mutational context alongside mutation co-occurrence to quantify the growth potential of variants within individuals. We found that gene-specific fitness differences can outweigh inter-individual variation and, therefore, could form the basis for personalized clinical management.

Age is the single largest factor underlying the onset of many cancers[1]. 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 multi-stage model of carcinogenesis[2]. However, the widespread presence of cancer-associated mutations in healthy tissues highlights the complexity of early detection and diagnosis of cancer[3–7].

CHIP is defined as the clonal expansion of HSPCs in healthy aged individuals. CHIP affects more than 10% of individuals over the age of 60 years and is associated with an estimated ten-fold increased risk for the later onset of hematological neoplasms[3–5]. There is a clear benefit of detecting CHIP early for close clinical monitoring and early detection, as the association between clone size and malignancy progression is well-established[5,8,9].

The particular mechanisms by which common mutations of CHIP—for example, *DNMT3A* and *TET2*—contribute to the progression of leukemia are still not understood, which hinders early diagnosis of CHIP on a gene or variant basis[8,10–12]. In clinical practice, CHIP is diagnosed by the presence of somatic mutations at variant allele frequencies (VAFs) of at least 2% in cancer-associated genes, that is in more than 4% of all blood cells[8,13]. 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[14,15]. As mutations in stem cells often drive leukemia[5], we hypothesized that stem cell fitness contributes substantially to transformation from CHIP to leukemia.

Stratification of individuals to inform close clinical monitoring for early detection or prevention of leukemia in the future will depend on the ability to accurately associate genes and their variants with progression to disease. However, it remains unresolved whether variant-specific or gene-specific fitness effects outweigh other factors contributing to variable progression among individuals, such as environment or genetics.

Hitherto, fitness effects have been predicted from large cross-sectional cohort data[14,16]. In this approach, single-timepoint data from many individuals are 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 timepoint creates additional uncertainty about whether a mutation has arisen recently and has grown rapidly (high fitness advantage) or arose a long time ago and has grown slowly (low fitness advantage). With longitudinal samples, fitness effects of individual mutations can be estimated directly from the change in VAF over multiple timepoints.

In this study, we worked with longitudinal data from the Lothian Birth Cohort of 1921 (LBC1921) and the Lothian Birth Cohort of 1936 (LBC1936)[17]. 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 of total timespan. We developed a new framework for extracting fitness effects from longitudinal data using Bayesian inference. First, a

[1]MRC Human Genetics Unit, University of Edinburgh, Edinburgh, UK. [2]Institute of Cancer Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. [3]Cancer Research UK Beatson Institute, Glasgow, UK. [4]Centre for Regenerative Medicine, Institute for Regeneration and Repair, University of Edinburgh, Edinburgh, UK. [5]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. [6]Edinburgh Clinical Research Facility, University of Edinburgh, Edinburgh, UK. [7]Paul O'Gorman Leukaemia Research Centre, Institute of Cancer Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. [8]Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, UK. [9]These authors contributed equally: Neil A. Robertson, Eric Latorre-Crespo, Linus J. Schumacher, Kristina Kirschner, Tamir Chandra. ✉e-mail:

likelihood-based filter for time series data (LiFT) allowed us to segregate between sequencing artifacts or naturally drifting populations of cells and fast-growing clones. Second, we inferred the growth potential or fitness effects simultaneously for all growing mutations within each individual and also allowed for clones with multiple mutations if these are favored by Bayesian model comparison. We detected gene-specific fitness effects within our cohorts, highlighting the potential for personalized clinical management.

Results

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 aging with approximately three-yearly follow-up for five waves, from the age of 70 years (LBC1936) and 79 years (LBC1921)[17]. We previously identified 73 participants with CHIP at wave 1 through whole-genome sequencing (WGS)[18]. Here, we used a targeted error-corrected sequencing approach using a 75-gene panel (ArcherDX/Invitae) to assess longitudinal changes in VAFs and clonal evolution over 21 years across both LBC cohorts (6 years in LBC1921 and 12 years in LBC1936; Supplementary Table 1). Error-corrected sequencing allowed accurate quantification, providing more sensitive clonal outgrowth estimates than our previous WGS data. We sequenced 248 LBC samples (85 individuals across 2–5 timepoints) and achieved a sequencing depth of 2,238× mean coverage (2,153× median) over all targeted sites with an average of 1.6 unique somatic variants (pan-cohort VAF 0.03–87%, median VAF 4.4%) detected per participant. We examined all participant-matched events across the time course: sequence quality control metrics revealed that only seven of 275 data points failed to meet our quality criteria, likely due to low initial VAF. Most of our variant loci generally displayed a high number of supporting reads, with a mean of 258 (Extended Data Fig. 1a).

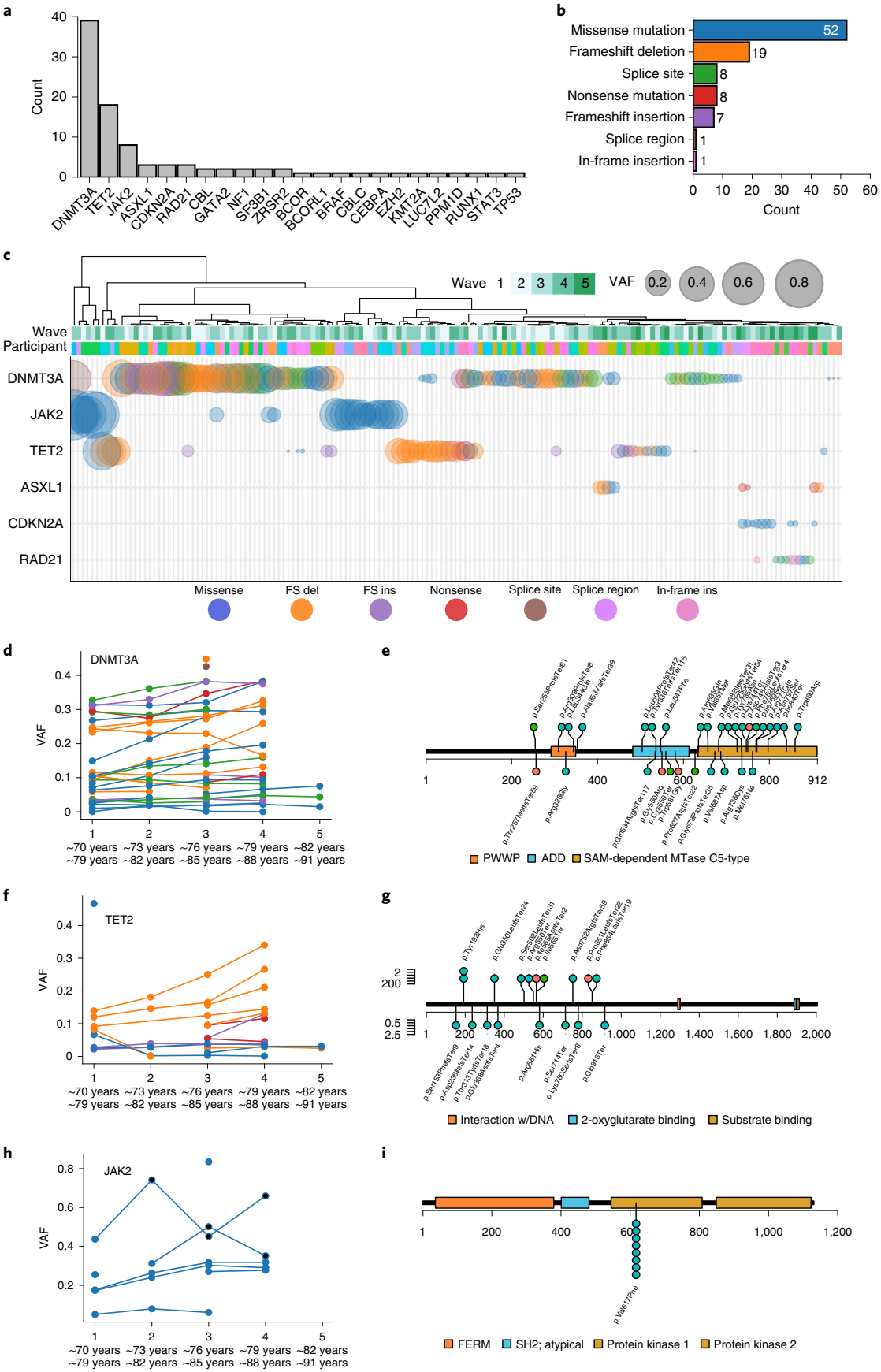
For our initial analysis, we retained variants with at least one timepoint at 2% VAF (Supplementary Table 2). *DNMT3A* was the most commonly mutated CHIP gene ($n=39$ events in 33 participants), followed by *TET2* ($n=18$ events in 15 participants), *JAK2* ($n=8$ events in eight participants) and *ASXL1* ($n=3$ events in three participants) (Fig. 1a–c and Extended Data Fig. 1e). Our mutation spectrum is consistent with previous studies in finding *DNMT3A* and *TET2* as the most frequently mutated genes[4,5]. We detected some variants more frequently at certain hotspots within a gene, such as R882H in *DNMT3A*, with previously unreported variants being present as well (Fig. 1d–i and Supplementary Table 2)[5]. We most frequently detected missense mutations with several other key protein-altering event types ranking highly, including frameshift insertions and deletions and nonsense mutations (Fig. 1a–c). Participants broadly cluster together across their time course, driven by the expanding or stable VAF of their harbored mutations, underscoring the high prevalence and large clone size of common clonal hematopoietic drivers, namely *DNMT3A*, *TET2* and *JAK2* (Fig. 1a–c). In the case of *JAK2V617F*, we identified two individuals who developed leukemia at wave 2 and received treatment between waves 2 and 3, likely driving a clear reduction in clone size (Fig. 1h). Those individuals were excluded from further analysis. In our data, we identified a lower frequency of mutations in splicing genes, such as *SF3B1*, despite the older age of the cohorts (Fig. 1a and Extended Data Fig. 1e). This is in contrast to previously published cohort data, where splicing mutations became more prominent with increased age[19]. Most mutations were missense, frameshift and nonsense mutations (Fig. 1b).

Overall, our sequencing approach allowed for high-resolution, longitudinal mapping of CHIP variants over 6-year and 12-year time spans in LBC1921 and LBC1936, respectively, and 21-year time span across both cohorts from the same geographical region and born 9 years apart.

Cataloguing of fitness effects for CHIP variants at >2% VAF. 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-specific 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[14], 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 and Extended Data Fig. 1e), we initially selected all CHIP variants in our data using the commonly used criterion of defining any variants with VAF > 2% as CHIP[8,13] and retaining only those variants with at least two timepoints (Fig. 2b). This approach identified 76 CHIP mutations overall (Fig. 2c). To estimate the fitness effect that each variant confers, we used Bayesian inference and birth–death models of clonal dynamics (Fig. 2a), including all trajectories with at least two timepoints (Supplementary Table 3). The resulting fitness values show an overall dependence of fitness on the gene level (Fig. 2d), with a wide distribution of fitness for some genes, such as *TET2* and *DNMT3A*, but not others, such as *JAK2* (which are all the same variant).

Longitudinal trajectories accurately stratify CHIP variants. Because 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 found that a VAF > 2% threshold not only misses fast-growing and potentially harmful variants (Fig. 2b) but can also include variants whose frequencies are shrinking (Fig. 2b,c) and, thus, either do not confer a fitness advantage or are being outcompeted by other clones. Overall, 70% of CHIP mutations detected by thresholding at 2% VAF were growing during the observed time span (Fig. 2b,c). Longitudinal data, thus, reveal limitations in defining CHIP mutations based on a widely used VAF threshold.

Fig. 1 | Clonal hematopoiesis in the LBCs. a, Counts of unique events that exceeded 2% VAF across the range of the longitudinal cohorts in our panel of 75 hematopoietic genes. **b**, Counts of the functional consequences of the unique events listed in Fig. 1a. Missense mutations, frameshift insertions and deletions and nonsense mutations are indicated. Exact counts, *n*, are for each category. **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 were clustered across all timepoints, with the point size scaled by VAF and colored by the functional consequence of the variant (as per Fig. 1b and legend). **d**, Clone size trajectories of all *DNMT3A* mutations across the time series in both LBC1921 and LBC1936 colored by the functional consequence of the variant (as per Fig. 1b,c). **e**, Locations of 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) with the structure of the gene labeled along the amino acid length of its protein. **f**. Clone size trajectories of all *TET2* mutations across the time series in both LBC1921 and LBC1936 colored by the functional consequence of the variant (as per Fig. 1b,c). **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 LBC1921 and LBC1936 colored by the functional consequence of the variant (as per Fig. 1b,c). Points marked in black denote timepoints after which the affected participant received treatment for leukemia. **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). All eight *JAK2* mutations are p.Val617Phe (*JAK2 V617F*) missense variants. del, deltion; FS, frameshift; ins, insertion.


a
Count
DNMT3A TET2 JAK2 ASXL1 CDKN2A RAD21 CBL GATA2 NF1 SF3B1 ZRSP2 BCOR BCORL1 BRAF CBLC CEBPA EZH2 KMT2A LUC7L2 PPM1D RUNX1 STAT3 TP53
b
Missense mutation 52
Frameshift deletion 19
Splice site 8
Nonsense mutation 8
Frameshift insertion 7
Splice region 1
In-frame insertion 1
Count
c
Wave 1 2 3 4 5
VAF 0.2 0.4 0.6 0.8
Wave
Participant
DNMT3A
JAK2
TET2
ASXL1
CDKN2A
RAD21
Missense FS del FS ins Nonsense Splice site Splice region In-frame ins
d
DNMT3A
VAF
~70 years ~73 years ~76 years ~79 years ~82 years
~79 years ~82 years ~85 years ~88 years ~91 years
e
PWWP ADD SAM-dependent MTase C5-type
f
TET2
VAF
~70 years ~73 years ~76 years ~79 years ~82 years
~79 years ~82 years ~85 years ~88 years ~91 years
g
Interaction w/DNA 2-oxyglutarate binding Substrate binding
h
JAK2
VAF
~70 years ~73 years ~76 years ~79 years ~82 years
~79 years ~82 years ~85 years ~88 years ~91 years
i
FERM SH2; atypical Protein kinase 1 Protein kinase 2

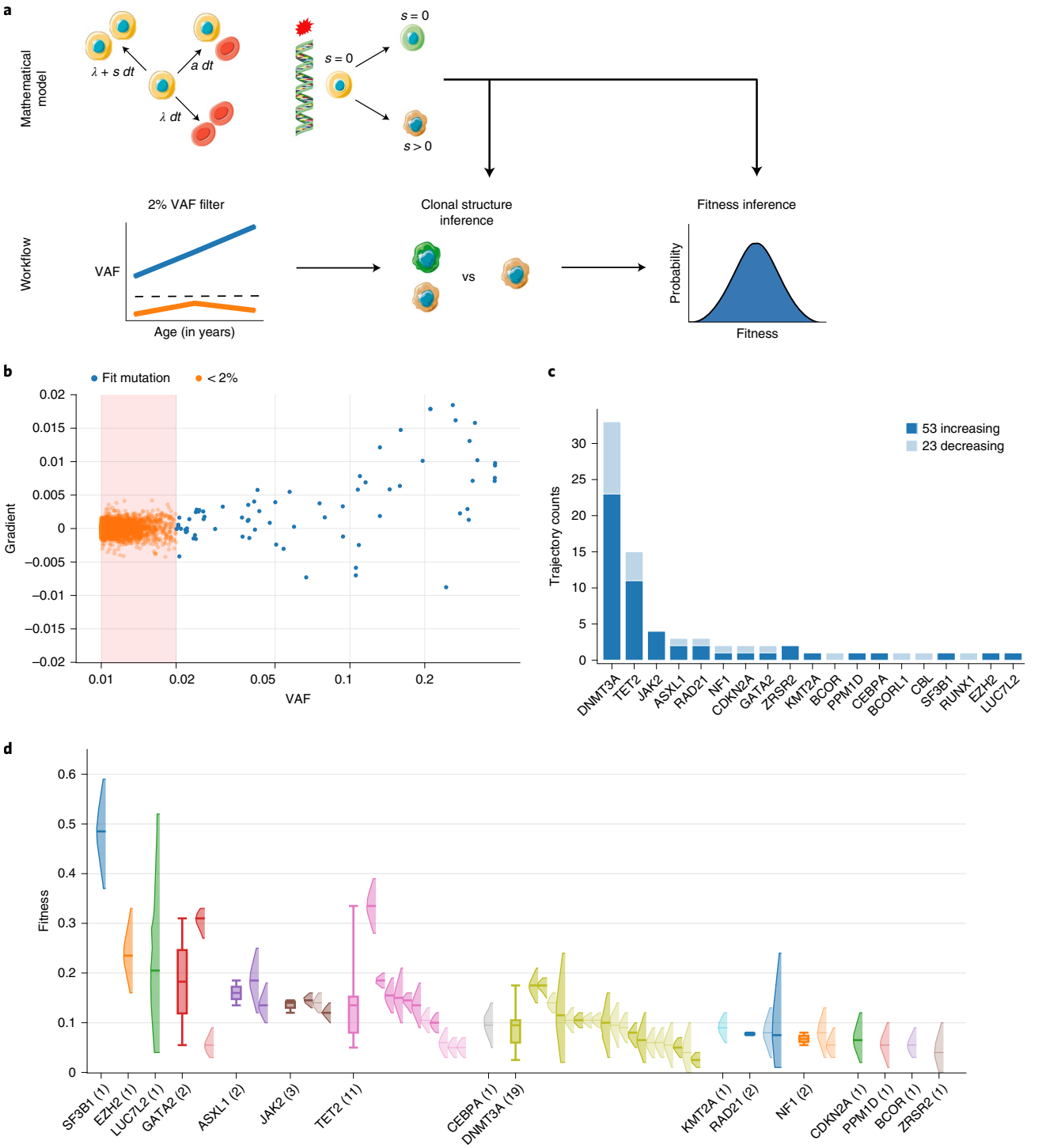


Fig. 2 | Fitness effects of variants at 2% VAF threshold in longitudinal data. a, Schematic of the mathematical model (top) and workflow (bottom) used to infer the fitness of mutations reaching VAF > 2% during the observed time span. Clonal structure and fitness inference are based on a mathematical model of clonal dynamics (Methods). HSPCs (top, yellow cells) naturally acquire mutations over time that can be neutral ($s = 0$, green cell) or increase self-renewal bias ($s > 0$, brown cell), leading to the formation of genetic clones. Artwork includes images by Servier Medical Art licensed under CC BY 3.0. **b**, VAF measurement $v(t_0)$ at initial timepoint t_0 versus gradient in VAF, $(v(t_{end}) - v(t_0))/(t_{end} - t_0)$, between initial and last timepoints t_0 and t_{end} of all variants detected in the LBCs with at least two timepoints. Each data point corresponds to a trajectory in the LBCs and has been colored according to its CHIP status based on the 2% VAF threshold (red box). Blue and orange, respectively, denote whether trajectories achieved a VAF > 2% during the observed time span or not. Note: VAF is displayed on a logarithmic scale, as most mutations are concentrated at low VAF. **c**, Number of trajectories passing the currently used 2% VAF threshold, broken down into whether VAF is increasing or decreasing from the first to last timepoint. **d**, Fitness effects of mutations grouped by gene and ranked by median fitness. The posterior probability distribution of the fitness as inferred from our model of clonal dynamics is displayed for each mutation (only the 90% interval is shown). The sample size, *n*, of observed variants in each gene is denoted in brackets. When more than one mutation is observed in a gene, we further display a box plot showing the median and exclusive interquartile range of the MAP fitness estimates associated with the gene.

To overcome the limitations of a threshold-based selection of fit variants, we sought to filter variants based on longitudinal information, by comparing a stochastic model of clonal dynamics with a model of sequencing artifacts (Fig. 3a). This novel approach, which we named LiFT, allows classification of fit variants even for VAF < 2%. LiFT classification of fit variants broadly agreed with noise profile statistics from the ArcherDX pipeline (Extended Data Fig. 2f,g) but identified additional variants by leveraging the longitudinal nature of the data. LiFT classification resulted in 114 variant trajectories (Fig. 3b–d and Extended Data Fig. 2a–g), 86% 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 2% VAF, with only 70% of variants identified to be growing and, thus, likely to confer a fitness advantage. Of the 114 variants we detected, 50 would not have been detected using the previous VAF threshold filter. We, therefore, recomputed fitness estimates for this new set of fit trajectories (Fig. 3e,f). Growing variants that were missed by the traditional filtering method include highly fit variants such as *U2AF1* Q157R (fitness 33.5%) and *DNMT3A* R882H (fitness 16%) (Fig. 3c,g and Supplementary Table 4). VAF thresholding did not identify any *TP53* variants. However, LiFT identified four *TP53* mutations, all of which were growing over the observed time course (Fig. 3c,g and Supplementary Table 4). In addition, all of those were either termination/frameshift mutations or previously reported as cancer-associated in the Catalogue of Somatic Mutations in Cancer (COSMIC)[20] and classified as likely damaging (Supplementary Table 5). Moreover, all *TP53* variants led to high fitness effects; thus, our filtering method allows us to identify potentially harmful variants at very low VAFs. Overall, the variants detected by LiFT were of higher fitness than those detected by VAF thresholding (Fig. 3f; Kruskal–Wallis $H = 14$, $P = 1 \times 10^{-4}$), with an even larger effect size when comparing variants that are exclusive to each filtering algorithm (Fig. 3f; Kruskal–Wallis $H = 18$, $P = 1 \times 10^{-5}$).

We further stratified variants using seven computational predictors recently identified as being most useful for identifying pathogenic mutations[21–27] (Fig. 3g and Supplementary Table 5), categorizing the most prevalent CHIP variants into likely damaging (21 variants), possibly damaging (20 variants) and likely benign (11 variants) as well as frameshifts and terminations (37 variants, which are also most likely damaging to protein structure and, thus, protein function; Supplementary Table 6). Our novel LiFT algorithm, therefore, produces a low false discovery rate of pathogenic variants, with 88% of the detected fit variants being predicted to be possibly damaging, frameshift or termination.

Taken together, applying a probabilistic model of clonal dynamics to longitudinal sequencing data results in a novel method—the LiFT algorithm—that improves on the threshold-based definition of CHIP mutations (Fig. 3a). The LiFT algorithm replaces an arbitrary cutoff on VAF by a choice of false discovery rate (through a Bayes factor threshold) and, as a result, selects fewer trajectories with shrinking VAF (Figs. 2b,c and 3b–d).

Clinical relevance of LiFT. We further analyzed 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. 4a,b). In particular, we found that mutations in *TP53*, *SF3B1* and *SRSF2* conferred a higher fitness advantage over mutations in commonly mutated CHIP genes, such as *JAK2* and *DNMT3A*. We also tested differences in fitness by genes when summarized into functional categories and found trajectories of genes involved in DNA methylation to have lower fitness than genes involved in splicing and genes for transcription factors that are relevant in development (Extended Data Fig. 3a,b).

Differences in the distribution of fitness allow us to predict the future growth of mutations from initial timepoints. For example, if a patient presents with a variant in a gene with 10% fitness at 1% VAF, its growth could be confidently measured after 7 months (Fig. 4c), warranting a clinical follow-up over that timeframe to confirm or revise the fitness estimate. Conversely, the time between observations places a lower bound on the fitness that can be measured for mutations of a given VAF (Fig. 4d). These data can then inform on the timeframe for close clinical monitoring and early detection of disease.

Ableson et al.[16] compared CHIP carriers who never developed acute myeloid leukemia (AML) with CHIP where individuals subsequently developed AML, and they found that the number of mutations, the mutational burden and the size of the larger driver clone were associated with the risk of progression to AML. In the present study, we carried out a survival analysis to correlate the maximum observed VAF of mutations and survival. This correlation was stronger in the older cohort (LBC1921) although not statistically significant (hazard ratio (HR) = 1.35; 95% confidence interval (CI) 0.83, 2.19; $P = 0.23$) due to the small sample size (Extended Data Fig. 3d and Supplementary Table 7). In the younger cohort (LBC1936), we found that survival better correlated with the speed of growth of a mutation, although this was, again, not statistically significant (HR = 1.35; 95% CI 0.76, 2.4; $P = 0.3$) (Extended Data Fig. 3d and Supplementary Table 7).

Notably, only two timepoints are necessary to apply LiFT, making this a widely applicable method for existing cohorts and future studies (Extended Data Fig. 3c). We propose the use of LiFT over thresholding for clinical practice.

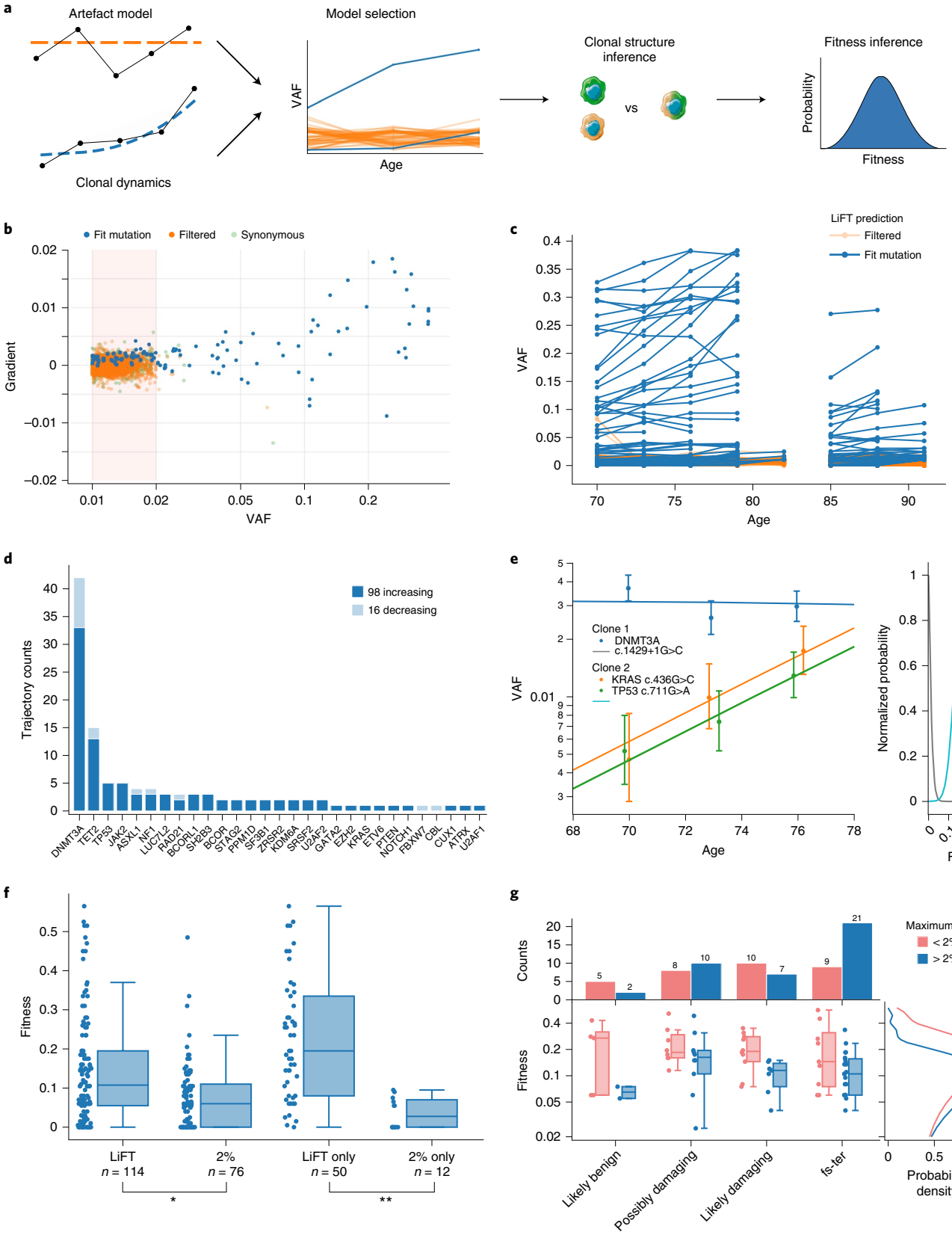
Discussion

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

Fig. 3 | LiFT allows classification of fit variants <2% VAF. a, Schematic of LiFT algorithm. LiFT compares a model of clonal dynamics (Fig. 1a) with an artifact model and performs Bayesian model selection. The subsequent steps to infer clonal structure and fitness distributions are as in Fig. 1a. **b**, Gradient in VAF versus VAF for variants detected in the LBCs with at least two timepoints and at least one VAF > 1% per trajectory, with filtered (orange), fit (blue) and synonymous (light green dots) mutations, classified by LiFT on a logarithmic scale. **c**, Longitudinal trajectories of fit (blue) and filtered (orange) mutations linked to age in years. **d**, Number of trajectories classified as fit by LiFT, broken down into increasing or decreasing VAF from first to last timepoint. **e**, Left, deterministic fit of all mutations selected by LiFT in an individual of the LBC cohorts using the inferred optimal clonal structure (Supplementary Information Methods, Appendix B). 90% CIs associated with binomial sampling noise are shown for each data point. VAF is displayed on a logarithmic scale. Right, posterior distribution of fitness associated to each clonal structure. **f**, Fitness effects of variants broken down by filtering method. The sample size, *n*, and statistical analyses comparing the distribution of fitness, computed using the non-parametric Kruskal–Wallis test, are highlighted (*$H = 14$, $P = 1 \times 10^{-4}$; **$H = 18$, $P = 1 \times 10^{-5}$). **g**, Fitness of variants selected as fit by LiFT broken down by their maximum VAF, >2% and <2%, and damage prediction. The top row displays a bar plot of variant counts for each category. The bottom row displays box plots showing the median and interquartile range of the distribution of MAP fitnesses by damaging prediction displayed on a logarithmic scale to emphasize relative differences in fitness between variants. Consequently, of a total of 89 variants with a damage prediction, 17 variants with fitness below 2% are not shown but are reported in Supplementary Tables 4–6. A marginal plot shows the Gaussian kernel density estimation of the MAP fitness values. fs, frameshifts; ter, terminations.

a gene basis and have shown considerable overlap in fitness coefficients among 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.

Our fitness estimates are independent of the time when the mutation was acquired. In cross-sectional studies, fitness



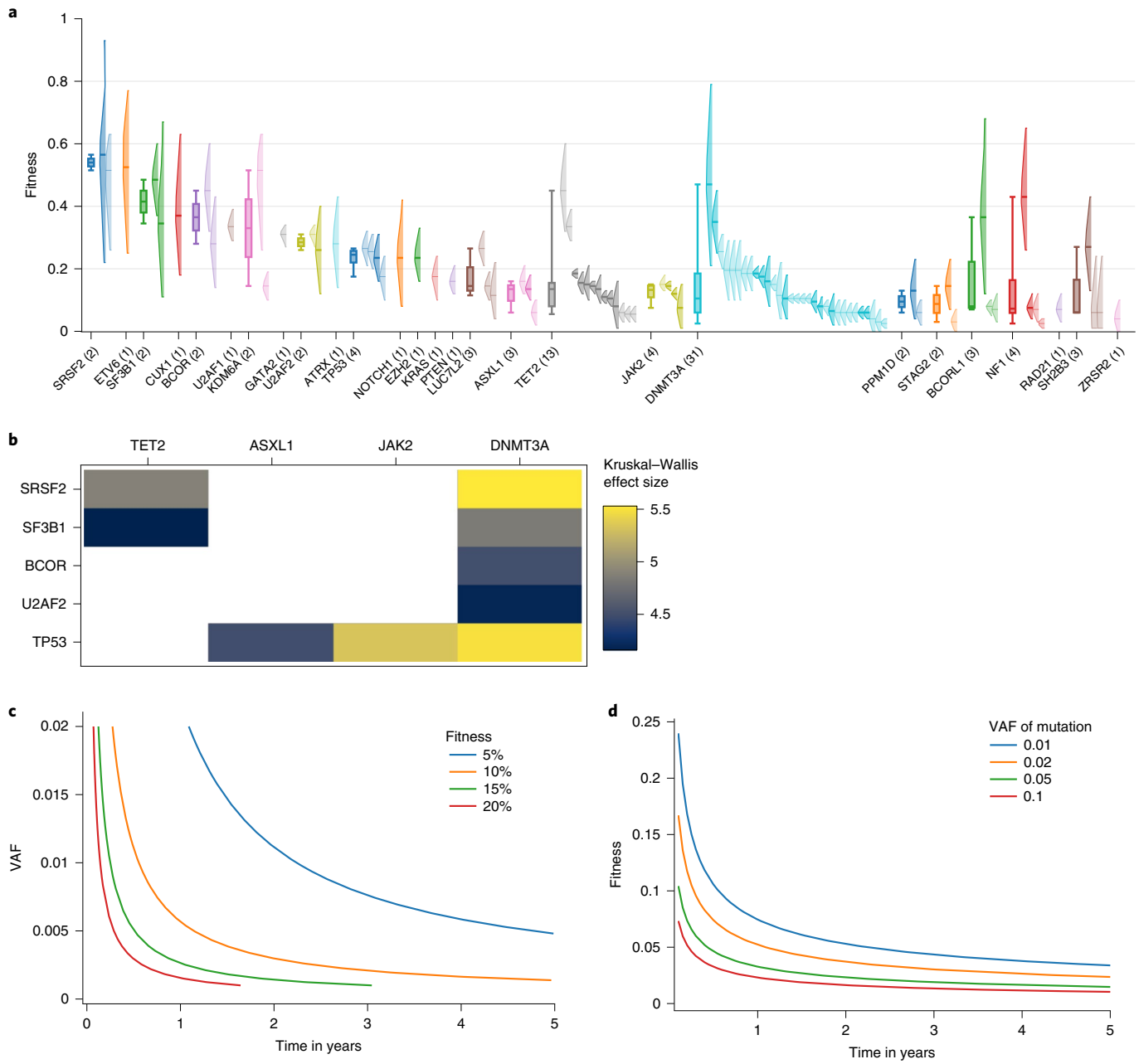


Fig. 4 | Clinical relevance of LiFT. a, Fitness effects of mutations selected as fit by the LiFT algorithm, grouped by gene and ranked by median fitness. The posterior probability distribution of the fitness as inferred from our model of clonal dynamics is displayed for each mutation (only the 90% interval is shown). The sample size, *n*, of observed variants in each gene is denoted in brackets. When more than one mutation is observed in a gene, we further display a box plot showing the median and exclusive interquartile range of the MAP estimates of fitness associated with the gene. **b**, Analysis of variance of the distribution of fitness across genes. Heat map of all statistically significant ($P < 0.05$) Kruskal–Wallis H statistics, labeled by effect size, computed for all combinations of pairs of genes. The effect size is only shown for statistically significant relations. Variants with a fitness below 2% were left out of this study, as our prediction classifies them as conferring no or a negligible fitness advantage. **c**, Minimum referral time in years based on 2 standard deviations below the expected growth of a clone given an initial VAF and fitness. Each line shows the initial size of mutation versus referral time for a given fitness. **d**, Minimum detectable fitness at referral observation based on 2 standard deviations below the expected growth of a clone given an initial VAF and fitness. Each line shows minimum detectable fitness versus referral time for an initial clone size.

estimates are generally (inversely) correlated with the mutation rate, introducing additional uncertainty[14]. In contrast, our fitness estimates are based on the observed growth among longitudinal samples and, thus, also take into account other mutations in an individual. The resulting fitness estimates are largely independent of hematopoietic stem cell absolute numbers (Extended Data Fig. 4b,c).

The strength of our approach, combining longitudinal data with our LiFT algorithm, is exemplified by *U2AF1* and *TP53*, for which no variants were identified by a 2% VAF threshold (Fig. 2b,c). In contrast, our LiFT method identified one *U2AF1* and four *TP53* variants, all of which are conferring a fitness advantage, scored as possibly damaging in our missense variant effect analysis and have been previously reported in COSMIC[20] (Fig. 3g and Supplementary

Tables 4 and 5). Moreover, we pick up the *DNMT3A R88H* variant with LiFT but not with 2% VAF thresholding—a mutation that is well-reported in the context of leukemia[28]. Therefore, for patients with those variants, close clinical monitoring for early detection of disease such as leukemia is merited.

Combining longitudinal data with LiFT enables a personalized approach managing CHIP (Extended Data Figs. 5 and 6). Longitudinal data allow quantifying fitness effects even for mutations not seen in large cohorts, as cross-sectional fitness estimation requires a mutation to be observed in multiple individuals. Our method offers clinicians a way forward for patient stratification even for unique variants occurring in single individuals, because two timepoints for one individual suffice to estimate fitness, including uncertainty quantification (Fig. 4e). We have provided a prediction of the time required between first and second observations to be able to accurately infer fitness, depending on the initial VAF of a mutation in an individual (Fig. 4c). For high fitness mutations (>10%), a follow-up clinical observation could be performed after only a few months, even for small clones (1% VAF or less). Conversely, the time between observations places a lower bound on the fitness that can be measured for mutations of a given VAF (Fig. 4d). In the future, these data can be used to inform time to the next appointment for close clinical monitoring of patients with clones containing highly fit variants, which will likely outcompete other clones. Using longitudinal data to better quantify and predict clonal progression in our study, however, comes with a tradeoff in the lower number of participants in our cohort and limits the power of cross-sectional analysis to find associations.

In addition, our inference method aims to resolve the clonal composition of multiple mutations in an individual. Specifically, we can now infer the likely co-occurrence of mutations from longitudinal data. Current cross-sectional studies do not take into account the clonal composition of individuals and, therefore, make predictions of the isolated effect of a mutation. In contrast, we are able to link fitness to clones carrying a specific combination of mutations that is unique to each individual, without relying on any prior knowledge of variant-specific fitness effects (Supplementary Table 4).

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/s41591-022-01883-3.

Received: 3 September 2021; Accepted: 27 May 2022;
Published online: 4 July 2022

References

1. de Magalhaes, J. P. How ageing processes influence cancer. *Nat. Rev. Cancer* **13**, 357–365 (2013).
2. Martincorena, I. Somatic mutation and clonal expansions in human tissues. *Genome Med.* **11**, 35 (2019).
3. Ayachi, S., Buscarlet, M. & Busque, L. 60 years of clonal hematopoiesis research: from X-chromosome inactivation studies to the identification of driver mutations. *Exp. Hematol.* **83**, 2–11 (2020).
4. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
5. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
6. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
7. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
8. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, eaan4673 (2019).
9. Park, S. J. & Bejar, R. Clonal hematopoiesis in cancer. *Exp. Hematol.* **83**, 105–112 (2020).
10. Terradas-Terradas, M., Robertson, N. A., Chandra, T. & Kirschner, K. Clonality in haematopoietic stem cell ageing. *Mech. Ageing Dev.* **189**, 111279 (2020).
11. Challen, G. A. & Goodell, M. A. Clonal hematopoiesis: mechanisms driving dominance of stem cell clones. *Blood* **136**, 1590–1598 (2020).
12. Shih, A. H., Abdel-Wahab, O., Patel, J. P. & Levine, R. L. The role of mutations in epigenetic regulators in myeloid malignancies. *Nat. Rev. Cancer* **12**, 599–612 (2012).
13. Steensma, D. P. & Bolton, K. L. What to tell your patient with clonal hematopoiesis and why: insights from two specialized clinics. *Blood* **136**, 1623–1631 (2020).
14. Watson, C. J. et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).
15. 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).
16. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
17. Taylor, A. M., Pattie, A. & Deary, I. J. Cohort profile update: the Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.* **47**, 1042–1042r (2018).
18. Robertson, N. A. et al. Age-related clonal haemopoiesis is associated with increased epigenetic age. *Curr. Biol.* **29**, R786–R787 (2019).
19. McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
20. Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
21. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics* **14**, S3 (2013).
22. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16**, S1 (2015).
23. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet* **99**, 877–885 (2016).
24. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020).
25. Raimondi, D. et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* **45**, W201–W206 (2017).
26. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
27. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
28. Ley, T. J. et al. *DNMT3A* mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363**, 2424–2433 (2010).

Methods

Participant samples and ethics. This study complies with all relevant ethical regulations. The study protocol was approved by NHS Lothian (formerly Lothian Health). Informed consent was given by all participants. Ethics permission for LBC1936 was obtained from the Multi-Centre Research Ethics Committee for Scotland (wave 1: MREC/01/0/56), the Lothian Research Ethics Committee (wave 1: LREC/2003/2/29) and the Scotland A Research Ethics Committee (waves 2, 3, 4 and 5: 07/MRE00/58). Ethics permission for LBC1921 was obtained from the Lothian Research Ethics Committee (wave 1: LREC/1998/4/183; wave 2: LREC/2003/7/23; wave 3: 1702/98/4/183) and the Scotland A Research Ethics Committee (waves 4 and 5: 10/MRE00/87).

LBC1921 contains a total of 550 participants at wave 1 of their testing (performed between 1999 and 2001) with a gender ratio of 234:316 (male:female) and a mean age at wave 1 of 79.1 years (s.d. = 0.6) (Supplementary Table 1)[17]. LBC1936 contains a total of 1,091 participants at wave 1 of their testing (performed between 2004 and 2007) with a gender ratio of 548:543 (male:female) and a mean age at wave 1 of 69.5 years (s.d. = 0.8) (Supplementary Table 1)[17]. We previously identified 73 participants with CHIP at wave 1 (ref. [18]). We sequenced DNA from those 73 LBC participants using a targeted gene panel (Supplementary Table 8) and added 16 LBC participants with previously unidentified CHIP and 4–5 timepoints. We have accepted 85 of 89 participants for inclusion in our study, removing four participants for failing to meet quality criteria (low library complexity), with a total of 248 samples together with 14 'Genome in a Bottle' (GIAB) controls, two per sequencing batch (Supplementary Table 9)[29]. In addition, two individuals carrying the *JAK2V617F* mutation received treatment for leukemia after the first respective timepoint available, potentially driving the observed reductions in clone size. Those patients were omitted from further analysis after sequencing (Fig. 1h).

Targeted, error-corrected sequencing and data filtering. DNA was extracted from Ethylenediaminetetraacetic acid (EDTA) whole blood using the Nucleon BACC3 kit (Sigma-Aldrich, GERPN8512), following the manufacturer's instructions. Libraries were prepared from 200 ng of each DNA sample using the Archer VariantPlex® 75 Myeloid gene panel and VariantPlex® Somatic Protocol for Illumina sequencing (Invitae, AB0108, and VariantPlex®-HGC Myeloid Kit for Illumina; Supplementary Table 9), including modifications for detecting low allele frequencies. Sequencing of each pool was performed using the NextSeq 500/550 High-Output version 2.5 (300 cycle) kit on the NextSeq 550 platform (Illumina). To inform reproducibility, background model for error and batch correction, we sequenced two GIAB DNA samples in each batch of samples (DNA NA12878, Coriell Institute)[29].

Reads were filtered for phred ≥30 and adapters removed using Trimmomatic (version 0.27)[30] before undergoing guided alignment to human genome assembly hg19 using bwa-mem (version 0.7.17)[31] and bowtie2 (version 2.2.1)[32]. Unique molecular barcodes (ligated before PCR amplification) were used for read de-duplication to support quantitative multiplexed analysis and confident mutation detection. Within targeted regions, variants were called using three tools (Lofreq (version 2.1.0)[33], Freebayes[34] and Vision (ArcherDX version 6.2.7, unpublished)), building a consensus from the output of all callers (Supplementary Table 2).

All filtered variants at 2% 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 in the Genome Aggregation Database (gnomAD; $P \leq 0.05$)[35]; (3) variants are not obviously germline variants (stable VAF across all waves ~0.5 or ~1) that may have been underrepresented in the gnomAD due to the narrow geographical origin of the LBC participants; and (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 artifact from the capture method (Supplementary Table 2). In addition, we manually curated this list, checking for variants that were previously reported, as per Jaiswal et al.[5], in COSMIC[20] or in the published literature (Supplementary Table 10). Finally, for any variant that surpassed the above criteria at VAF ≥ 2% across the measured time period, we included other participant-matched data points regardless of VAF level (Extended Data Fig. 1a,b).

To further mitigate against the diverse sources of noise that can occur in any sequencing experiment, which can become especially problematic when attempting to detect variants at low VAFs, the ArcherDX variant-calling platform leverages the pan-dataset coverage levels of each sample and the GIAB controls to establish a position-specific noise profile and, thus, ascertain the limit of detection (LOD) for each variant discovered in our panel. Here, we report two parameters for each variant: (1) the minimal detectable allele fraction (95% MDAF; Extended Data Fig. 1c), which describes the minimum VAF that a variant can be detected in our data, in essence describing the LOD for each event; and (2) the VAF outlier *P* value, which denotes the probability that any variant call could have been generated by sequencing noise given the position-specific noise distribution across our GIAB controls and the pan-dataset coverage levels of our samples, thus allowing us to discern overrepresented sequencing artifacts from real events (Extended Data Fig. 1d).

Computational prediction of missense variant effects. To predict which missense variants are most likely to be damaging, we used seven computational variant effect predictors recently identified as being most useful for identifying pathogenic mutations[21–27]. 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 version 2.1 for each computational predictor. We then averaged these fractions across all predictors. Note that DeepSequence[26] was not run for all proteins due to its computational intensiveness and difficulty of running on long protein sequences. We also performed predictions of missense variant (de)stabilization using FoldX 5.0, using the experimentally determined protein structure, if available, and the AlphaFold model[36,37].

Mathematical model of clonal dynamics to infer fitness. Given the longitudinal nature of this study, we can use the probabilistic solution of an established minimal model of cell division[14,38] to infer the parameter distribution resulting in the observed time evolution of VAF trajectories in a participant's genetic profile (Fig. 2a). For each individual, we simultaneously estimated the fitness of variants as well as the size of the stem cell pool, without needing to estimate the time of mutation acquisition.

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 VAF 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 $N(t) = N_w + \sum_{i\in I} n_i(t)$, where N_w is the average number of wild-type HSPCs in the individual. The bias toward self-renewal of symmetric divisions is parameterized by parameter s and determines the fitness advantage of a clone. In normal hematopoiesis, $s = 0$, in which case clones undergo neutral drift. For clones with non-neutral (fitness-increasing) mutations, $s > 0$, and this average clone size grows exponentially in time as $e^{s(t-t_0)}$ from an initial population of one SC at the time of mutation acquisition t_0. The full distribution of clone sizes is well-approximated by a negative binomial distribution matching the mean (exponential growth) and variance of the full stochastic solution (Supplementary Information Methods, section 1, and Extended Data Fig. 4a). Because the model dynamics are Markovian (without memory), once we condition on a previously observed timepoint in a trajectory, the prediction for all future times is independent of t_0. From the predicted clone size distributions, we can infer the marginal posterior distribution of parameter s using Bayes' theorem (Supplementary Information Methods, section 3)[39]. We further take into account the sampling error during sequencing to estimate the distribution of clone sizes at the start and end of each time interval in the longitudinal sequencing data. Here, we approximate this sampling error as binomial.

When multiple fit clones are present in an individual, we constrain the inference to share the SC pool size $N(t)$ for all variant trajectories in this individual. This increases the data:parameter ratio and produces richer dynamics, where the evolution of exponentially growing clones can be suppressed by the growth of a fitter clone. 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.

We take into account possible clonal substructures for all fit variants in an individual, selecting models with co-occurring mutations on the same clone if they are more likely after biasing against models with multiple mutations per clone, as these are presumed to be rarer (Supplementary Information Methods, section 2.4.7). The evidence supporting the optimal clonal structure, determined by Bayesian model comparison, relative to the model assuming no mutations co-occur on the same clone is shown in Extended Data Fig. 4d. We then infer the posterior fitness distributions per clone for the most likely clonal model in every participant.

Once we have inferred the posterior distributions of the parameters, we use the mode of the distribution (maximum a posteriori (MAP) estimate) for each mutation to visualize the deterministic—that is, average—growth curves. These result in the logistic time evolution of its corresponding VAF,

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

where we determine the time of mutation acquisition t_0, which is used only for plotting, using maximum likelihood (Supplementary Information Methods, Appendix B). Although deterministic fits are not a direct reflection of the inference results of our stochastic model, these can be used to visually assess the 'goodness of fit' of the fitness MAP estimates and have been included for each participant in LBC1921 and LBC1936, respectively, in Extended Data Figs. 5 and 6.

Note that this model cannot account for loss-of-heterozygosity events.

LiFT. To select fit variants, we compare the likelihood of the clonal model, including binomial sampling error, to a model of sequencing artifacts. The artifact

model assumes that all variability arises from sampling error with a proportion that remains constant over time. For variants that occur more than once in our dataset, we use a beta-binomial model to account for overdispersion, and, for unique variants, we use a binomial model. We select variants as fit only if the model evidence for the clonal model is at least four times that of the artifact model (Supplementary Information Methods, section 2.4, and Extended Data Fig. 2c,d). Fit variants thus selected are taken through to clonal structure model selection and fitness inference as described above.

Workflow overview. A workflow chart describing the full pipeline and implementation guidance is included in the GitHub repository (see 'Code availability'). Our pipeline can be applied to other datasets with a few adjustments. Our LiFT algorithm has been tailored to the LBC dataset by extracting parameters from the distribution of synonymous mutation reads, which inform the priors used for our Bayesian inference method (Supplementary Information Methods, section 2.3.3, and Extended Data Fig. 2a–c). Guidance on how to adapt our LiFT algorithm to other datasets is included in the code repository. All other parts of the pipeline, including the extraction of variants using ArcherDx software and the inference of clonal structures and fitness, are directly applicable to other datasets.

Framework implementation. Both LiFT and Bayesian inference of the posterior distribution of model parameters were implemented in Python version 3.7 (ref. [40]) with dependencies on Numpy version 1.21.5 (ref. [41]), Scipy version 1.7.3 (ref. [42]) and Pandas version 1.3.4. Survival analysis was implemented using Python version 3.7 (ref. [40]) with dependencies on lifelines version 0.26.4 (ref. [43]). Data curation was undertaken in Python version 3.7 (ref. [40]) and R base[44], with use of the 'tidyverse'[45] suite of packages and plotted with ggplot2 (ref. [46]).

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

Data availability

We have deposited all data pertinent to this analysis, including the de-identified raw FASTQ read data and processed variant calls for our longitudinal cohort, onto the National Center of Biotechnology Information Gene Expression Omnibus under accession ID GSE178936. LBC phenotypic data are available in the database of Genotypes and Phenotypes (dbGAP) under accession number phs000821.v1.p1. All other Lothian Birth Cohort data are deposited in dbGAP or are provided via the LBC Data Access Collaboration (https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration). Information concerning the cohort is contained here, including its history, data summary tables for both LBC1921 and LBC1936 and data access request forms and contact information to obtain all data points (contact: https://www.ed.ac.uk/profile/simon-cox, ; timeframe: 1 month to respond).

Code availability

All code used in this manuscript is available at https://github.com/neilrobertson/LBC_ARCHER.

References

29. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
32. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
33. Wilm, A. et al. 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 (2012).
34. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
35. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
36. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
37. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
38. Till, J. E., Mcculloch, E. A. & Siminovitch, L. A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proc. Natl Acad. Sci. USA* **51**, 29–36 (1964).
39. Bayes, T. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. https://doi.org/10.1098/rstl.1763.0053 (1763).
40. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. https://dl.acm.org/doi/book/10.5555/1593511 (CreateSpace, 2009).
41. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
42. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
43. Davidson-Pilon, C. lifelines: survival analysis in Python. *J. Open Source Softw.* **4**, 1317 (2019).
44. R Core Team. R: a language and environment for statistical computing. https://www.R-project.org/ (R Foundation for Statistical Computing, 2021).
45. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
46. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. https://ggplot2.tidyverse.org (Springer-Verlag, 2016).

Acknowledgements

We gratefully acknowledge the contributions of the LBC participants and members of the LBC research team who collect and manage the LBC data. We thank C. P. Ponting for critical reading of the manuscript. We would also like to thank B. Tait for his help, advice and patience throughout. LBC1921 was supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC) (SR176) to I.J.D; by a Royal Society–Wolfson Research Merit Award to I.J.D.; and by the Chief Scientist Office of the Scottish Government's Health Directorates (CZB/4/505; ETM/55) to I.J.D. LBC1936 is supported by the BBSRC and the Economic and Social Research Council (BB/W008793/1) to S.R.C; Age UK (Disconnected Mind project, which supports S.E.H) to I.J.D and S.R.C.; the Medical Research Council (MR/M01311/1 to I.J.D. and MR/K026992/1 to S.R.C.); and the University of Edinburgh. K.K. is funded by a John Goldman Fellowship, sponsored by Leukaemia U.K. (2019/JGF/003 to K.K.) and received CRUK Glasgow Centre funding (C7932/A25142 to K.K.) and CRUK Scotland Centre funding (CTRQQR-2021\100006 to K.K.). M.T.T. and N.A.R. are supported by Medical Research Council-funded Ph.D. studentships (MR/N013166/1 to M.T.T. and N.A.R.). T.C. and L.S. are supported by Chancellor's Fellowships held at the University of Edinburgh. J.A.M. is a Lister Institute Research Prize Fellow. E.L.C. is a cross-disciplinary postdoctoral fellow supported by funding from the University of Edinburgh and the Medical Research Council (MC_UU_00009/2). S.R.C. is supported by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and the Royal Society (221890/Z/20/Z). We are also grateful for funding from the Howat Foundation (grant holder, M.C.).

Author contributions

L.J.S., K.K. and T.C. conceived and supervised the study. N.A.R., E.L.C., L.J.S., K.K. and T.C. wrote the manuscript. L.M., A.F. and L.M.G. generated data. N.A.R. and E.L.C. developed the methodology for data analysis. N.A.R., E.L.C., M.T.T., A.C.P., J.A.M., B.J.L., J.L.P., R.F.H., R.E.M. and J.L.P. conducted data analysis. S.E.H., S.R.C. and I.J.D. curated the LBCs and gave access to samples. M.C. advised on aspects of the study.

Competing interests

K.K. received a reagent grant from ArcherDX/Invitae. L.M. consults for Illumina. M.C. has received research funding from Cyclacel and Incyte, is/has been an advisory board member for Novartis, Incyte, Pfizer and Jazz Pharmaceuticals and has received honoraria from Astellas, Novartis, Incyte, Pfizer and Jazz Pharmaceuticals. The remaining authors declare no competing interests.

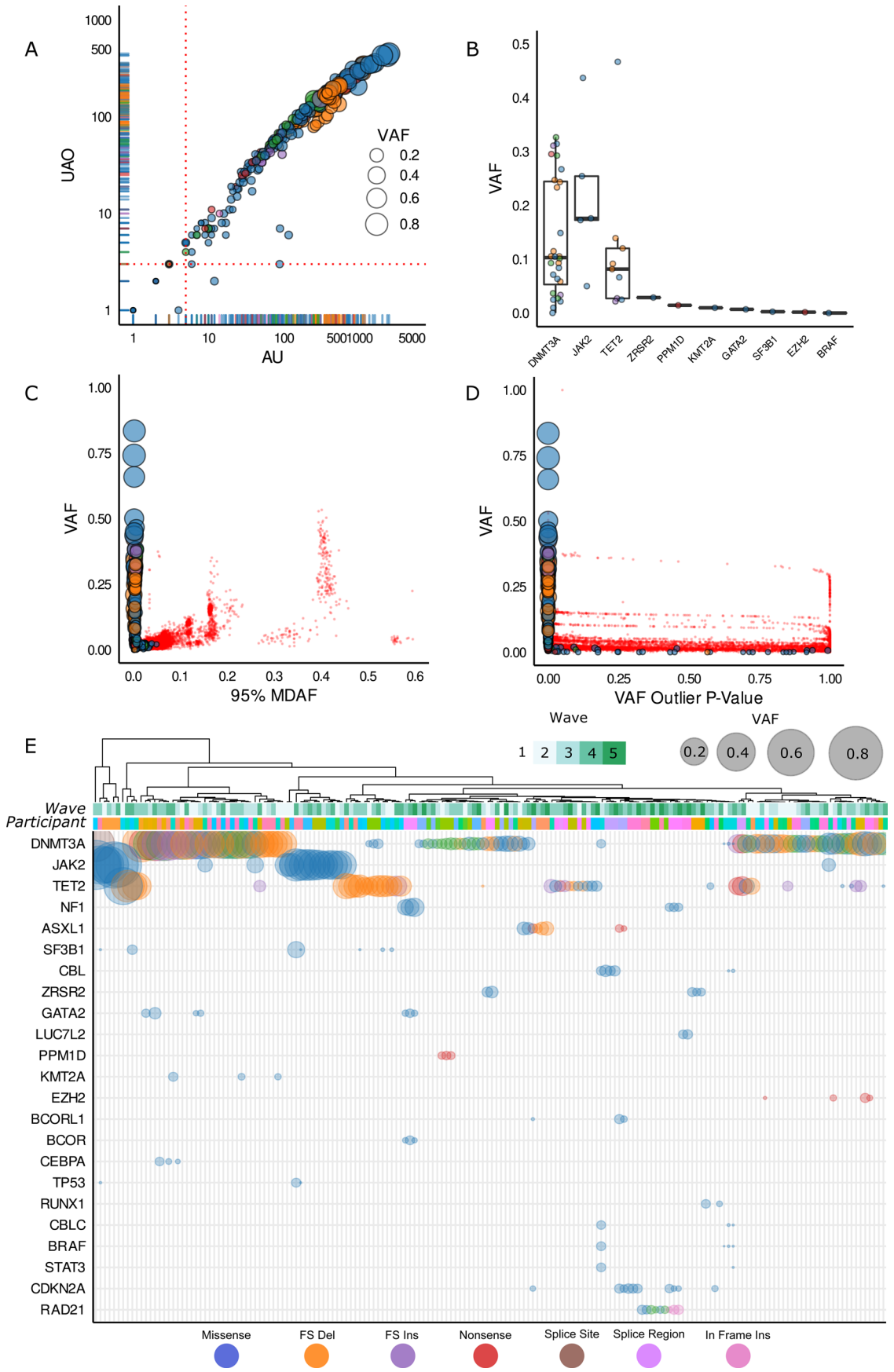
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-022-01883-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-01883-3.

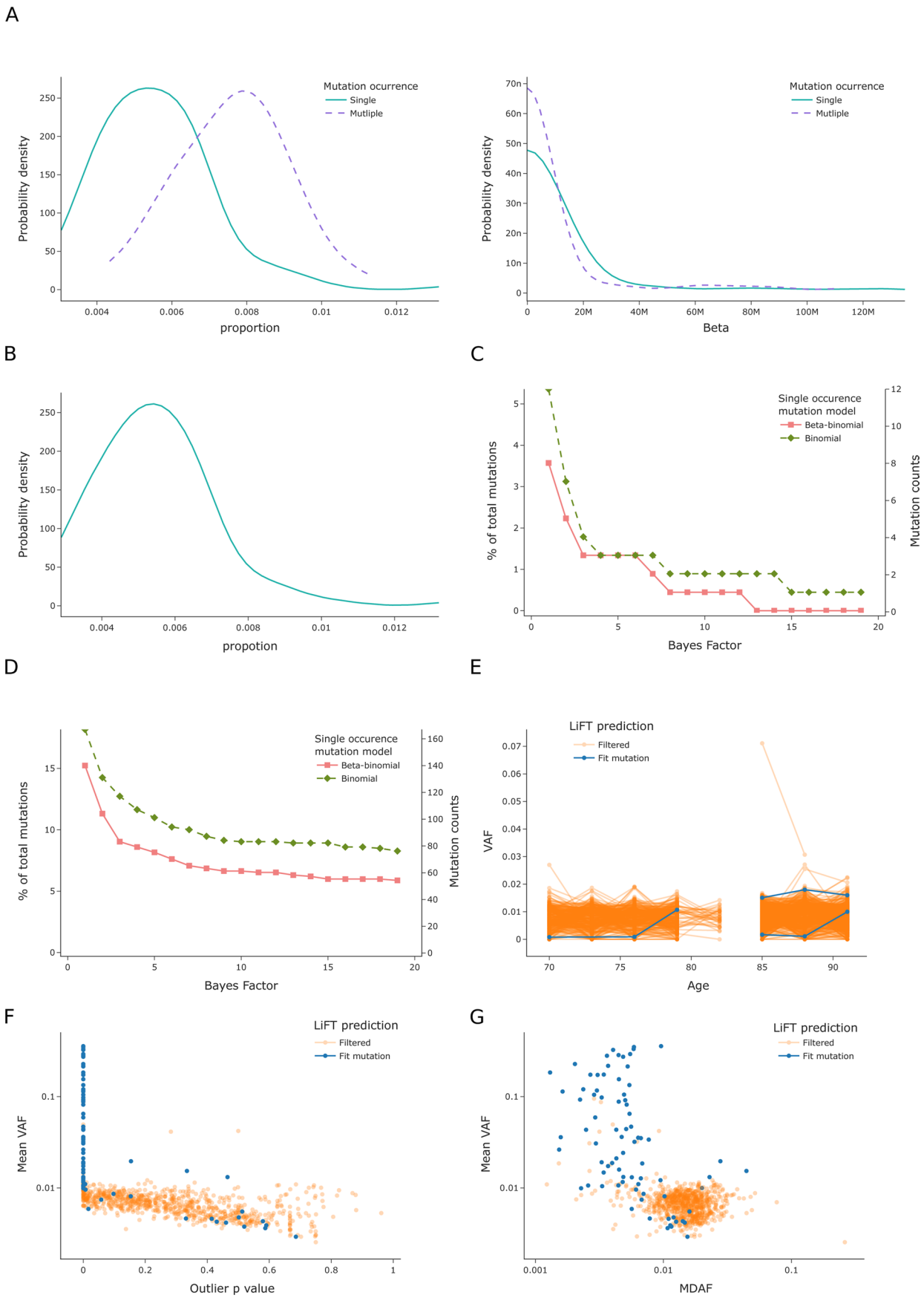
Correspondence and requests for materials should be addressed to Linus J. Schumacher, Kristina Kirschner or Tamir Chandra.

Peer review information *Nature Medicine* thanks Jamie Blundell, Alejo Rodriguez-Fraticelli, Hubert Serve and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary handling editor: Anna Maria Ranzoni, in collaboration with the *Nature Medicine* team.



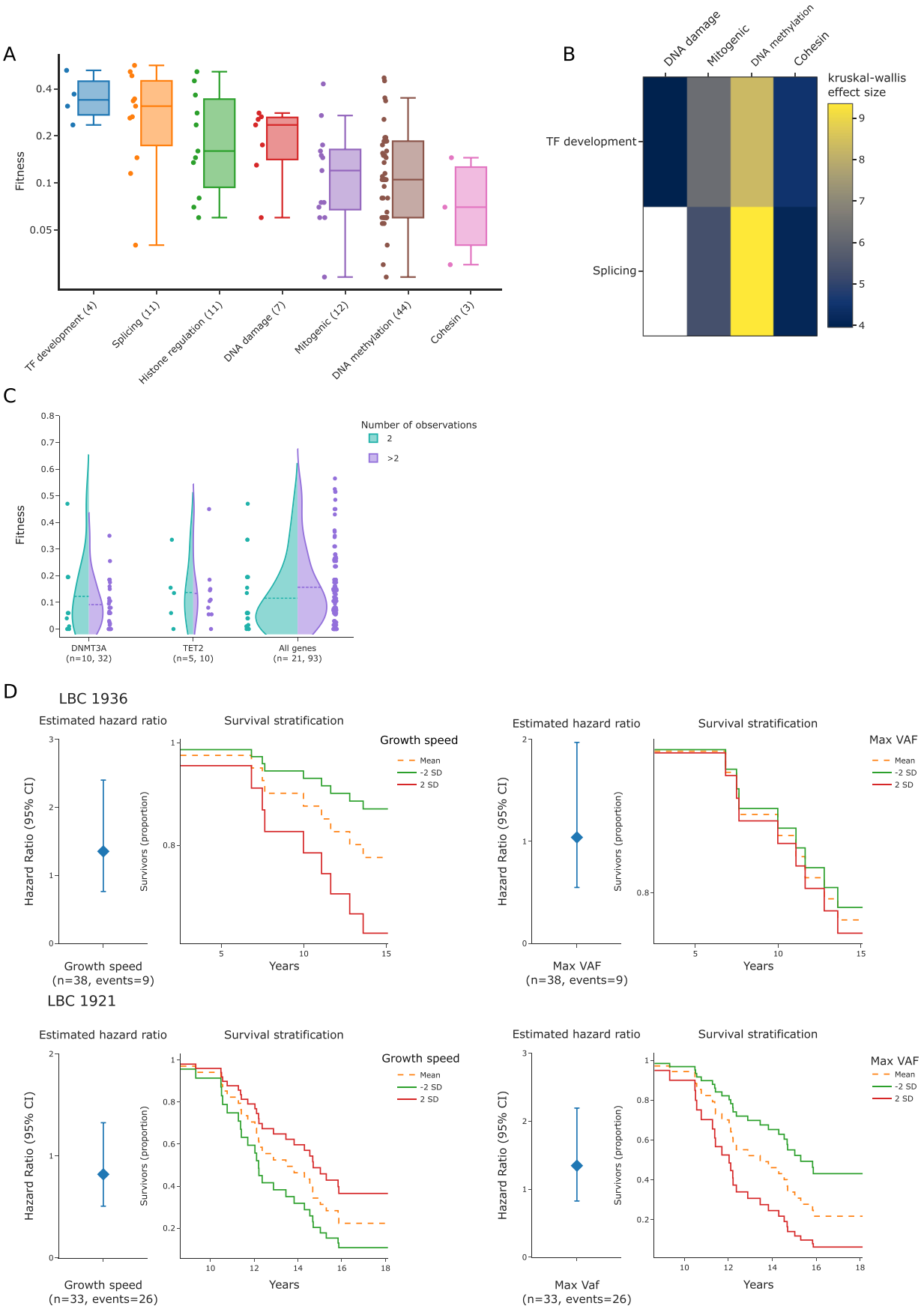
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Quality Control Metrics. a. Sequence quality metrics for mutation calls across participants and time-points filtered for 2% VAF. Plotted 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 - a measure of molecular complexity). Red dotted lines denote filter thresholds in both measurements ($AO \geq 5$, $UAO \geq 3$) and points are scaled by the VAF of the somatic mutation. Only 7 (of 275) data points failed to meet our filter criteria which were not excluded 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 2% VAF coloured by variant classification and ordered by largest mean VAF showing the median and interquartile range. **c**. The 95% MDAF (Minimal Detectable Allele Fraction with 95% Confidence) versus the VAF for each event. All variants used in our analysis above 2% VAF are scaled by their clone size and coloured by their functional consequence. Points in red are events that failed to pass our quality criteria and are removed from subsequent work. **d**. The VAF Outlier P-Value (describing the pan-cohort position-specific background noise) versus VAF for each event. All variants used in the analysis above 2% VAF are scaled by their clone size and coloured by their functional consequence. Points in red are events that failed to pass our quality criteria and are removed from subsequent work. All accepted events that exceed VAF Outlier P-Value > 0.1 are generally low VAF and are supported by matching events across the time-series that adhere to our acceptance criteria of VAF Outlier P-Value ≤ 0.1. **e**. Schematic of all affected genes in the cohort with the largest clone size of an event in any given gene shown above 2% VAF. 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).



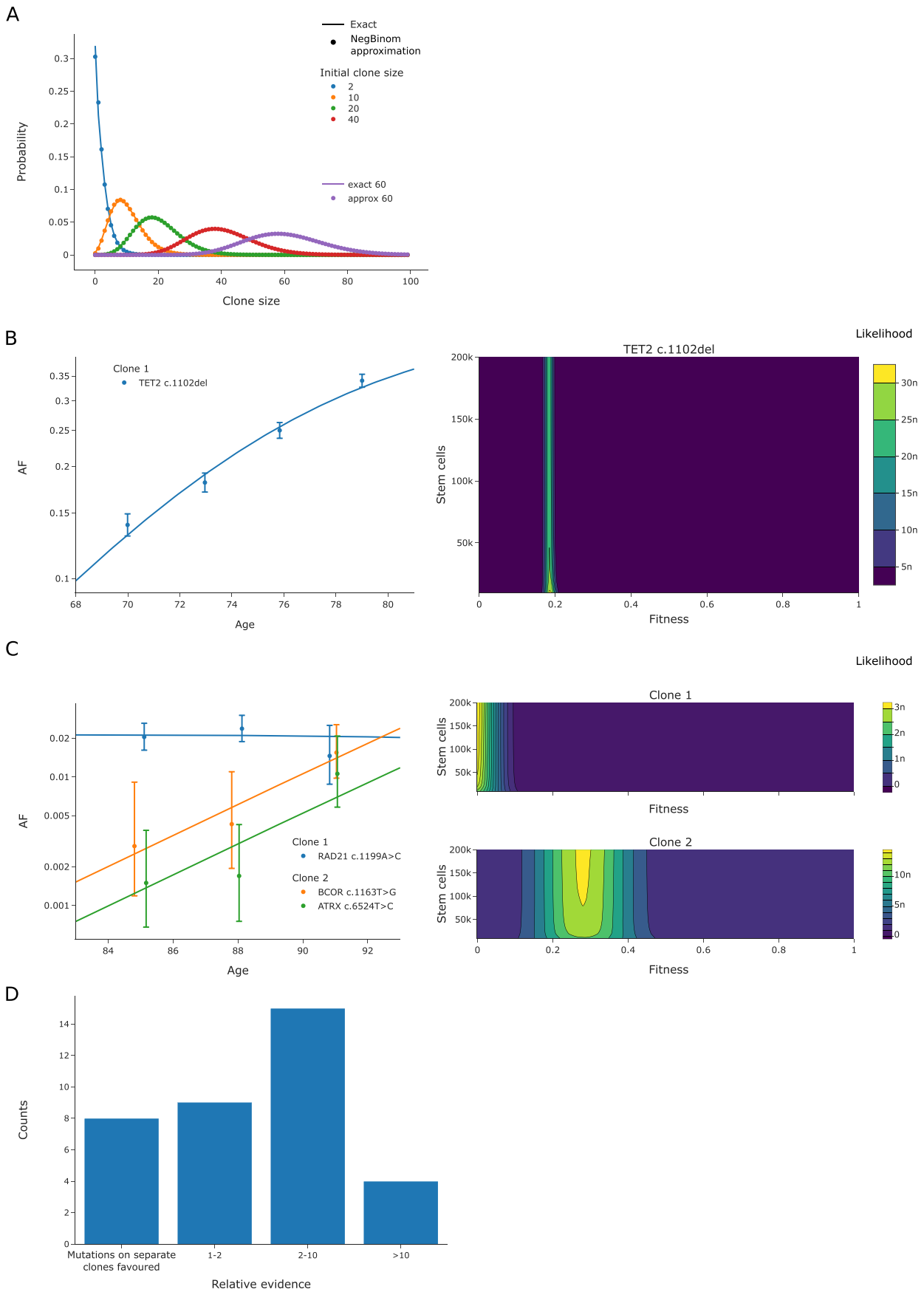
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | LiFT Method Details. a. Prior distributions for the beta-binomial model for sequencing artefacts. Priors are constructed separately for mutations with a single occurrence and mutations with multiple observations in the LBCs (see SI methods Section 2.3). **b**. Prior distribution of the proportion for the binomial model for sequencing artefacts. This prior is constructed only for mutations with a single occurrence in the LBC. **c**. Effect of the Bayes Factor (BF) threshold on the number of non-synonymous variants selected as fit using LiFT. In red, we show the results assuming that sequencing artefacts always follow a beta-binomial model, regardless of the mutation occurrence in the LBC. In green, we show the results where the sequencing artefact model assumes a binomial model for single occurring mutations and a beta-binomial model for mutations with multiple occurrences in the LBCs. **d**. Effect of the BF on the number of synonymous variants selected as fit using LiFT. Colour coding as in Fig. S2C. **e**. Longitudinal trajectories of non-synonymous variants coloured by their LiFT status; fit (blue) and filtered (orange). **f**. Comparison between LiFT status and the VAF Outlier P-value. Each data point corresponds to a trajectory in the LBC and has been coloured according to its LiFT status; fit (blue) and filtered (orange). The coordinates of each data point are given by the average VAF Outlier p-Value and their average VAF. **g**. Comparison between LiFT status and the Minimal Detectable Allele Fraction (MDAF). Each data point corresponds to a trajectory in the LBC and has been coloured according to its LiFT status; fit (blue) and filtered (orange). The coordinates of each data point are given by the average MDAF and their average VAF. Note that the MDAF is shown on a logarithmic scale.



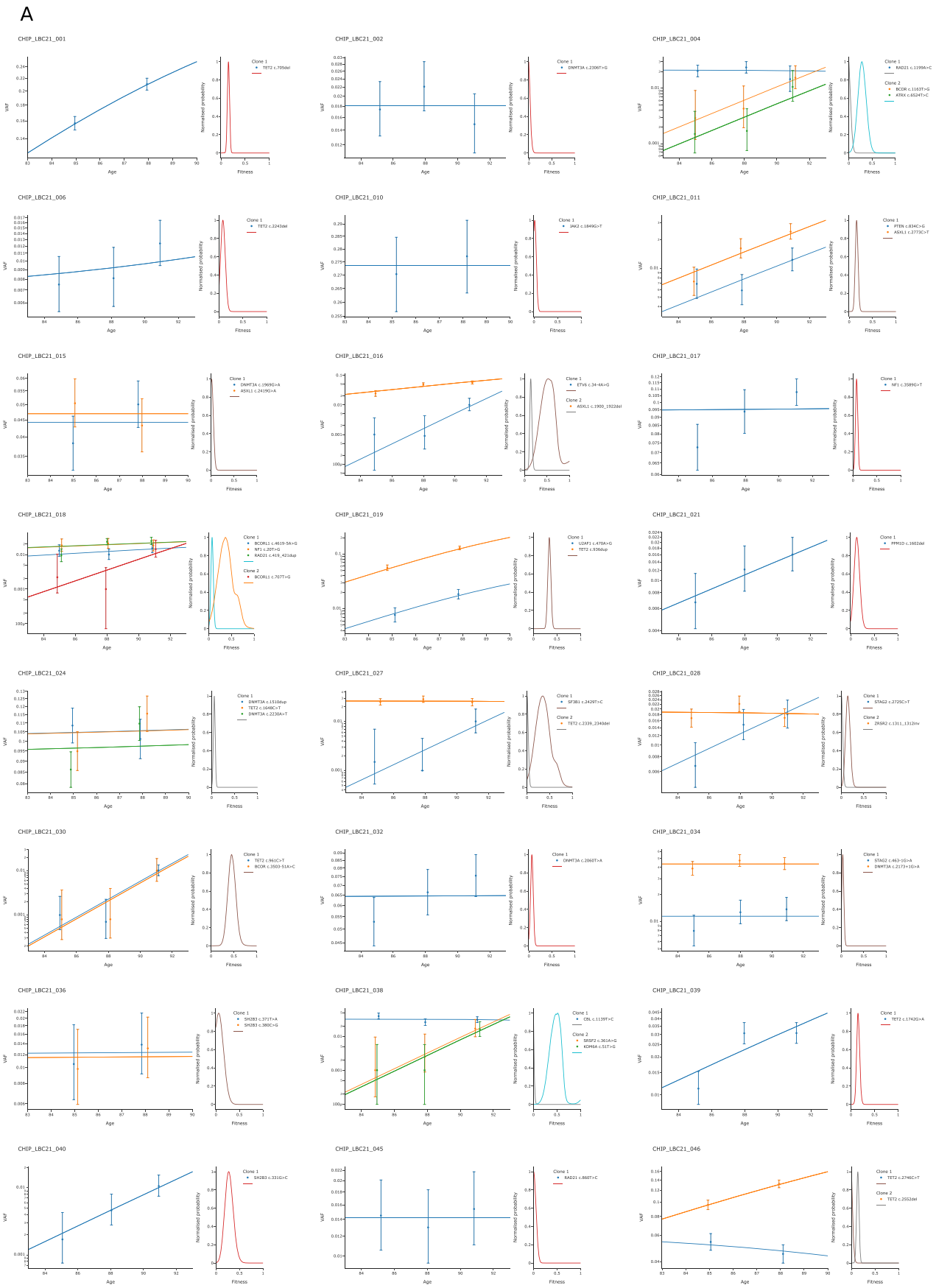
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Clinical Relevance of LiFT - Supporting Material. a. Distribution of fitness by gene category. Genes are grouped according to their biological function into DNA methylation (*TET2, DNMT3A*), Splicing (*SF3B1, U2AF1, SRSF2, U2AF2, ZRSR2, LUC7L2, DDX41*), mitogenic function (*KRAS, NF1, JAK2, JAK3, SH2B3, PTEN, PTPN11, NRAS*), cohesin (*RAD21, STAG2*), DNA damage (*TP53, CDKN2A, PPM1D, ATRX*) and Transcription factors (TF) important during development (*GATA2, RUNX1, NOTCH1, CUX1, ETV6*). The sample size, n, of each gene category is denoted in brackets. For each gene category we display a boxplot showing the maximum a posteriori (MAP) estimates of fitness for variants in the category, as well as the median and exclusive interquartile range. **b**. Analysis of variance of the maximum posterior fitness estimates across gene categories. Heatmap of all statistically significant ($p < 0.05$) Kruskal-Wallis H statistics, labelled by effect size, computed for all combinations of pairs of genes. The effect size is only shown for statistically significant relations. Variants with a fitness below 2% were left out of this study as our prediction classifies them as conferring no or a negligible fitness advantage. **c**. Influence of the number of time-points in a trajectory on the inferred fitness distributions. We show the maximum posterior estimates for genes *DNMT3A* and *TET2* and for all LiFT variants split according to the number of time-points. **d**. Survival analysis (Cox proportional hazards regression model) broken down by cohort and covariates. LBC1921 and LBC1936 are analysed separately given their difference in age during the observed time-span. (left) Error bar showing the inferred hazard ratio coefficient and 95% CI for each regression study, as well as the sample size, n, and the number of observed events in each analysis. Note that none of the survival analyses shown are statistically significant. The complete summary for each analysis is found in Supplementary Table 7. (right) Kaplan-Meyer survival plots for the LBC cohort stratified using 2 standard deviations of the analysed covariate.



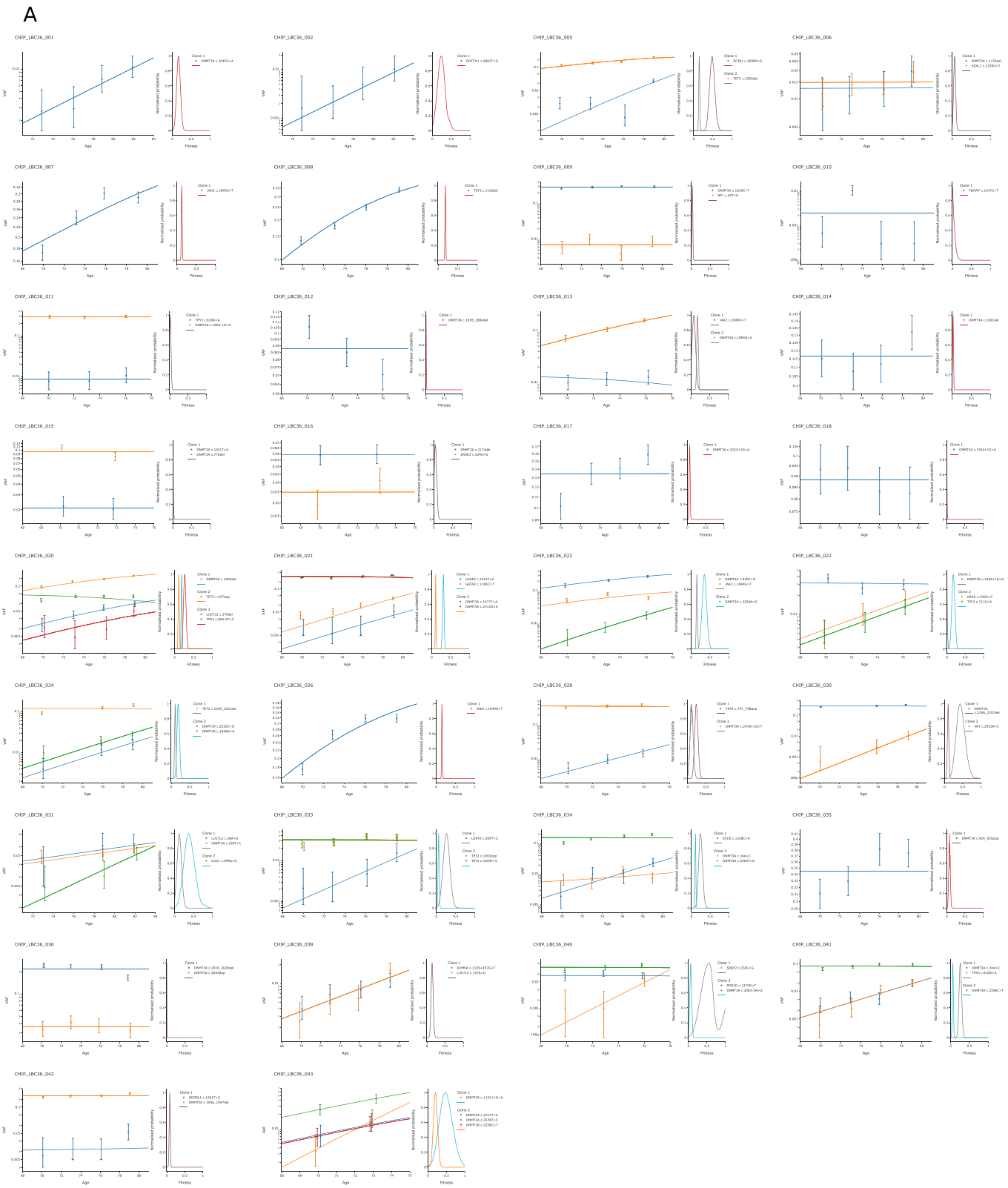
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Clonal Dynamics and Inference - Supporting Material. . **a**. Approximation of a neutral birth-death model using the negative binomial distribution. The exact model assumes symmetric divisions occur every 40 weeks, or 1.3 divisions per year, and has no bias towards self-renewal (see SI methods Section 1). **b**. Deterministic trajectory (see SI methods Appendix B) with maximum a posteriori (MAP) fitness and fitted time of mutation (left) and joint posterior distribution of fitness and number of wild-type HSPCs population (right) inferred from an individual with a single mutation selected by LiFT. 90% confidence intervals associated with binomial sampling noise are shown for each data point. Note that VAF is displayed on a logarithmic scale to highlight relative differences and the initial exponential growth of clones. Also note that a small random horizontal jitter has been added to data points to avoid overlapping of confidence intervals. **c**. Deterministic trajectory (see SI methods Appendix B) with maximum posterior fitness and fitted time of mutation (left) and joint posterior distribution of fitnesses and number of wild-type HSPCs inferred from an individual with three mutations, selected by LiFT, occurring in two clones. 90% confidence intervals associated with binomial sampling noise are shown for each data point. Note that VAF is displayed on a logarithmic scale to highlight relative differences and the initial exponential growth of clones. Also note that a small random horizontal jitter has been added to data points to avoid overlapping of confidence intervals. **d**. Evidence supporting the clonal structure selected by our Bayesian model comparison relative to the model assuming no mutations co-occur on the same clone. The evidence is only shown for non-trivial cases where more than one mutation was selected by LiFT in an individual.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Deterministic Visualisation of Mutational Trajectories in the LBC21. a. Deterministic trajectories (see SI methods Appendix B) with maximum a posteriori (MAP) fitness and wild-type stem cells and fitted time of mutation (left) and posterior distribution of fitness associated to each clonal structure (right) inferred for all mutations selected by LiFT in each participant of the LBC1921 cohort. 90% confidence intervals associated with binomial sampling noise are shown for each data point. Note that VAF is displayed on a logarithmic scale to highlight relative differences and the initial exponential growth of clones. To use a logarithmic axis, data points with zero observations have been replaced by VAF = 0.001, or a factor of 10 below our observation threshold. Also note that a small random horizontal jitter has been added to data points to avoid overlapping of confidence intervals.



Extended Data Fig. 6 | Deterministic Visualisation of Mutational Trajectories in the LBC36. a. Deterministic trajectories (see SI methods Appendix B) with maximum a posteriori (MAP) fitness and wild-type stem cells and fitted time of mutation (left) and posterior distribution of fitness associated to each clonal structure (right) inferred for all mutations selected by LiFT in each participant of the LBC1936 cohort. 90% confidence intervals associated with binomial sampling noise are shown for each data point. Note that VAF is displayed on a logarithmic scale to highlight relative differences and the initial exponential growth of clones. To use a logarithmic axis, data points with zero observations have been replaced by VAF = 0.001, or a factor of 10 below our observation threshold. Also note that a small random horizontal jitter has been added to data points to avoid overlapping of confidence intervals.

nature research

Corresponding author(s): Linus Schumacher, Kristina Kirschner, Tamir Chandra

Last updated by author(s): Jan 5, 2022

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	No software was used for data collection
Data analysis	All the code developed for the data analysis of this article is publicly available and documented in the Github repository: https://github.com/neilrobertson/LBC_ARCHER Workflow overview A workflow chart describing the full pipeline and implementation guidance are included in the github repository (see Code Availability). Our pipeline can be applied to other datasets with a few adjustments. Our LiFT algorithm has been tailored to the LBC dataset by extracting parameters from the distribution of synonymous mutation reads which inform the priors used for our Bayesian inference method (see SI methods Section 2.3.3 and Extended Data Fig. A, B and C). Guidance on how to adapt our LiFT algorithm to other datasets is included in the code repository. All other parts of the pipeline, including the extraction of variants using ArcherDx software and the inference of clonal structures and fitness, are directly applicable to other datasets. Data analysis is split in 3 parts: 1) Variant calling, data aggregation and quality control * Variant calling and raw data processing was completed using the ArcherDX/Invitea analysis pipeline. This was received as a virtual machine that was hosted within the University of Edinburgh (Virtualisation Team). https://analysis.archerdx.com/static/Archer_Analysis_Manual_4_1_0.pdf * Data aggregation was performed using python and R scripts contained in a sub-folder within the Git repository. Data curation was undertaken in Python v.3.7 and R base with use of the "tidyverse" suite of packages and plotted with ggplot2.

2) Longitudinal analysis of variants.
Longitudinal analysis of variants is implemented in Python v.3.7 with dependencies on Numpy v.1.21.5, Scipy v.1.7.3 and Pandas 1.3.4.
Survival analysis was implemented using Python v.3.7 with dependencies on Lifelines 0.26.4.

3) Survival analysis.
Survival analysis is implemented in Python programming language using Lifelines package.

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

We have deposited all data pertinent to this analysis including the de-identified raw fastq read data and processed variant calls for our longitudinal cohort onto the NCBI Gene Expression Omnibus (Geo) with accession ID: GSE178936 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178936). LBC phenotypic data are available at dbGAP under the accession number phs000821.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000821.v1.p1).

All other Lothian Birth Cohort Data are deposited in dbGAP or provided via the LBC DAC (https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration). Information concerning the cohort is contained here: including its history, data summary tables for both LBC1921 and LBC1936 with data access request forms and contact information to obtain all data points.

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

Life sciences study design

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

Sample size	Sample size was not estimated but samples were selected on the basis of harbouring CHIP variants in previous whole genome sequencing (WGS) at wave one. Our methodology is designed to determine fitness estimates in single samples, ergo, sample size was deemed sufficient.
Data exclusions	The study was initially focused on healthy cognitive ageing; at the inception of the study (at approximately age 70), participants were recruited if they reported no dementia or other neurodegenerative diagnoses. In addition, to take part in the first wave of the study, participants had to have been born in 1936 in Scotland, and be living in the Edinburgh and Lothians area of Scotland when recruited. Some data-points
Replication	not applicable
Randomization	not applicable
Blinding	not applicable

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	The Lothian Birth Cohort 1921 (LBC1921) contains a total of 550 healthy 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). The Lothian Birth Cohort 1936 (LBC1936), contains a total of 1091 healthy 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) (Taylor et al., 2018)). Both cohorts are Scottish cohorts. Participant characteristics of the whole cohort are described in the articles cited in the box directly below. Participants characteristics of the specific subsample used in the present study are described in the Methods section of the manuscript. There is known range restriction among members of LBC1921 and LBC1936. They were healthier and better educated than members of the general population of the same age, e.g. Johnson et al 2011 Health Psychology 30:1-11.
Recruitment	Recruitment for LBC1921 was similar to LBC1936. The Lothian Birth Cohort 1921 participants were identified for invitation to participate via newspaper advertisements and also via linkage with the NHS to identify addresses of those individuals born in 1921 living in the region (most schoolchildren in 1932 had taken the Scottish Mental Survey 1932 at school, and the study was designed as a follow-up of those older adults, aged ~79 at recruitment). The Lothian Birth Cohort 1936 participants were identified for invitation to participate via newspaper advertisements and also via linkage with the NHS to identify addresses of those individuals born in 1936 living in the region (most schoolchildren in 1936 had taken the Scottish Mental Survey 1947 at school, and the study was designed as a follow-up of those older adults, aged ~70 at recruitment. The study protocol papers that describe this recruitment process - along with the ethical approvals - are described in detail in the following open access protocol papers: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2222601/ https://pubmed.ncbi.nlm.nih.gov/22253310/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6124629/
Ethics oversight	Ethics permission for the Lothian Birth Cohort 1936 was obtained from the Multi-Centre Research Ethics Committee for Scotland (Wave 1: MREC/01/0/56), the Lothian Research Ethics Committee (Wave 1: LREC/2003/2/29), and the Scotland A Research Ethics Committee (Waves 2, 3, 4 & 5: 07/MRE00/58). Ethics permission for the Lothian Birth Cohort 1921 (LBC1921) was obtained from the Lothian Research Ethics Committee (Wave 1: LREC/1998/4/183; Wave 2: LREC/2003/7/23; Wave 3: 1702/98/4/183) and the Scotland A Research Ethics Committee (Waves 4 and 5: 10/MRE00/87).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Appendix 6: Manuscript – Clonality in Haematopoietic Stem Cell Ageing (2020)

Mechanisms of Ageing and Development 189 (2020) 111279

Contents lists available at ScienceDirect

Mechanisms of Ageing and Development

journal homepage: www.elsevier.com/locate/mechagedev

Clonality in haematopoietic stem cell ageing

Maria Terradas-Terradas[a], Neil A. Robertson[b], Tamir Chandra[b,*], Kristina Kirschner[a,*]

[a] Institute of Cancer Sciences, University of Glasgow, Glasgow G61 1BD, UK
[b] MRC Human Genetics Unit, University of Edinburgh, Edinburgh, EH4 2XU, UK

ARTICLE INFO

Keywords:
Clonal haematopoiesis of indeterminate potential
Environment
Cell-Intrinsic
Ageing
DNMT3A
TET2

ABSTRACT

Clonal haematopoiesis of indeterminate potential (CHIP) is widespread in the elderly. CHIP is driven by somatic mutations in leukaemia driver genes, such as Janus Kinase 2 (*JAK2*), Tet methylcytosine dioxygenase 2 (*TET2*), ASXL Transcriptional Regulator 1 (*ASXL1*) and DNA (cytosine-5)-methyltransferase 3A (*DNMT3A*), leading to reduced diversity of the blood pool. CHIP carries an increased risk for leukaemia and cardiovascular disease. Apart from mutations driving CHIP, environmental factors such as chemokines and cytokines have been implicated in age-dependent multimorbidities associated with CHIP. However, the mechanism of CHIP onset and the relationship with environmental and cell-intrinsic factors remain poorly understood. Here we contrast cell-intrinsic and environmental factors involved in CHIP development and disease propagation.

1. Introduction

Age is the single most significant factor underlying the onset of many haematological malignancies (de Magalhães, 2013), with changes in the clonal composition towards a myeloid bias commonly occurring with advanced age (Cho et al., 2008). The onset of clonal haematopoiesis of indeterminate potential (CHIP) in the haematopoietic stem and progenitor cell (HSPC) compartment is also associated with haematological malignancies (Genovese et al., 2014). CHIP is apparent in the general population from age 60 with a steady increase in prevalence to 18–20% of individuals aged over 90 years at 2% variant allele frequency (VAF) (McKerrell et al., 2015). CHIP is driven by somatic mutations in leukaemic driver genes, thereby reducing the diversity of the stem cell pool. Epigenetic modifiers such as *TET2*, *ASXL1* and *DNMT3A* are the most frequently mutated genes in CHIP. *TET2* and *DNMT3A* are epigenetic regulators involved in DNA methylation impacting self-renewal and differentiation capacities of haematopoietic stem cells (HSCs) while *ASXL1* - a member of the polycomb repressive complex - is involved in chromatin remodelling and affects hematopoietic repopulating capacity and expansion of the haematopoietic stem cell compartment (Bowman et al., 2018; Challen et al., 2011; Fuster et al., 2017; Jeong et al., 2018; Ko et al., 2011; Li et al., 2011; Lu et al., 2016; Mayle et al., 2015; Moran-Crusio et al., 2011; Pan et al., 2017; Quivoron et al., 2011; Wang et al., 2014). Interestingly, one hallmark of ageing is the global loss of methylation and profound changes to heterochromatin (Chandra and Kirschner, 2016; Cypris et al., 2019).

JAK2V617F is a common synonymous variant that is frequently mutated in CHIP and age-related myeloid malignancies (Chen et al., 2012), where the *JAK2* tyrosine phosphatase is constitutively activated driving a plethora of downstream pathways such as the phosphoinositide-3-kinase/Protein kinase B pathway (PI3K/AKT), Signal Transducers and Activators of Transcription (STAT) and RAS/RAF/MEK/ERK Mitogen-activated protein kinase pathways. Together these pathways confer a proliferative advantage, resistance to DNA damage mediated apoptosis and can activate an inflammatory response (Chen et al., 2012).

DNA damage response and stress-related genes Tumour Suppressor 53 (*TP53*) and Protein Phosphatase, Mg2+/Mn2+ Dependent 1D (*PPM1D*) are another class of mutations identified in CHIP (Genovese et al., 2014; McKerrell et al., 2015). *TP53* and *PPM1D* mutations are predominantly mutated in leukocytes of patients who have undergone cancer treatment for solid tumours and display clonal haematopoiesis (Coombs et al., 2017), associating genotoxic stress with clonal selection. In this study, clonal haematopoiesis was associated with secondary leukaemia development following solid cancer therapy. In other studies elucidating the mechanism of mutant *PPM1D* in clonal haematopoiesis and subsequent therapy-related myeloid malignancies (Hsu et al., 2018; Kahn et al., 2018), mutant *PPM1D* seemed to confer resistance to apoptosis in the context of genotoxic stress. Therapy-induced senescence is a prominent feature in cancer therapy and an alternative mechanism of cancer therapy resistance, with *TP53* being one of the most prominent senescence inducers, engaging a specific, downstream senescence programme that differs profoundly from apoptosis (Kahlem et al., 2004; Kirschner et al., 2015). Whether *TP53* mutations shift the

* Corresponding authors.
E-mail addresses: (T. Chandra), (K. Kirschner).

https://doi.org/10.1016/j.mad.2020.111279
Received 5 March 2020; Received in revised form 24 April 2020; Accepted 1 June 2020
Available online 08 June 2020

pathway away from apoptosis towards senescence in the context of CHIP remains to be elucidated.

Lastly, splicing factors emerge late in the pathogenesis of CHIP with the most prominent being Splicing Factor 3b Subunit (*SF3B1*), *SRSF2* (Serine and Arginine Rich Splicing Factor 2) and U2 Small Nuclear RNA Auxiliary Factor 1 (*U2AF1*) (Inoue et al., 2016). These mutations alter RNA splicing in a sequence-specific manner and might affect downstream pathways leading to CHIP over a longer period of time.

CHIP is associated with an increased risk for haematological cancers and all-cause mortality, specifically coronary heart disease and ischaemic stroke, for which age is a major risk factor (Genovese et al., 2014; Jaiswal et al., 2017, 2014; McKerrell et al., 2015; Zink et al., 2017). In addition, CHIP in patients with heart failure results in increased mortality and hypertension (Dorsheimer et al., 2019). In this review, we will discuss factors that can lead to the onset of CHIP and its age-associated diseases contrasting fitness acquired by mutations (cell-intrinsic) against cell-extrinsic changes in the ageing environment.

2. Cell-intrinsic contributions to CHIP

Several landmark studies reported the occurrence of CHIP in healthy aged individuals using various deep sequencing approaches of peripheral blood mononuclear cells (PBMCs). The major driver mutations, such as *JAK2*, *TET2* and *DNMT3A,* in all cohorts examined overlap, with varying allele frequencies of the distinct driver mutations. A few genes were only reported in some cohorts, such as the DNA damage response pathway gene *PPM1D* (Genovese et al., 2014; Jaiswal et al., 2017, 2014; McKerrell et al., 2015). The occurrence of CHIP mutations with age pointed to a time-dependent acquisition pattern, leading to a competitive advantage and driving CHIP with advanced age (Fig. 1). This is especially true when considering splicing factors.

Mouse studies of commonly mutated CHIP genes support the notion that clonal outgrowth is driven by cell-intrinsic properties in the HSPC compartment. Among the most frequently mutated genes is *Dnmt3a*. Challen and colleagues demonstrated that conditional loss of *Dnmt3a* in HSCs impairs their differential potential by altering DNA methylation (Challen et al., 2011). In knockout mice, loss of *Dnmt3a* immortalises HSCs (Jeong et al., 2018), leading to skewed division potential with HSCs being primed towards self-renewal for up to twelve rounds of transplantation with gradual and focal loss of DNA methylation at key HSC self-renewal sites. In this study, however, transformation required additional mutations (Jeong et al., 2018). In a study by Mayle et al., transplantation of murine *Dnmt3a*-knockout HSCs into irradiated wild-type mice resulted in the development of a range of haematological malignancies, leading to increased mortality. These results suggest a cell-intrinsic role of *Dnmt3a* loss in HSCs in acquiring a preleukemic state (Mayle et al., 2015). This is in accordance with another mouse study, where HSPCs with a *DNMT3aR882H* mutation, the most commonly found CHIP mutation, promotes leukaemia only in the presence of other oncogenes such as N-RasG12D (Lu et al., 2016). *DNMT3aR882H* alone led to hypomethylation at cis-elements of essential stemness genes such as the Meis Homeobox 1 (*Meis1)*, MN1 proto-oncogene (*Mn1*), and Hoxa gene clusters, and led to increased expression of a panel of stemness genes (Lu et al., 2016).

TET2 is also commonly mutated in individuals with CHIP. Several studies using *Tet2* knockout and mutant mouse models have explored cell-intrinsic mechanisms of HSCs leading to malignant transformation (Ko et al., 2011; Moran-Crusio et al., 2011; Quivoron et al., 2011). All these studies showed an expansion of the HSC compartment, as well as enhanced self-renewal potential associated with TET2 loss-of-function. Interestingly, Li and colleagues demonstrated an increase of the murine HSC pool in *Tet2* knockout mice, with only a subset of mice developing myeloid malignancies (Li et al., 2011). A later study showed spontaneous development of different haematological malignancies in *Tet2* knockout mice, resulting from increased mutagenicity (Pan et al., 2017). Single-cell -targeted sequencing revealed a higher mutation rate

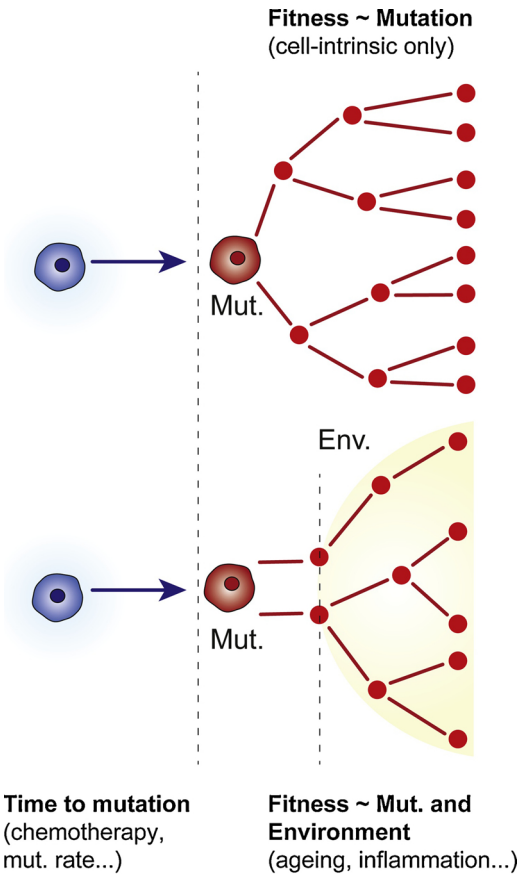


Fig. 1. Is CHIP dependent on the environment or driven by cell intrinsic factors?
Upper panel: **CHIP is driven cell-intrinsically.** Here, the time to acquisition of the CHIP mutation (Mut.) and the change in fitness conferred by the mutation are the determining factors of clonal outgrowth. Average time to mutation depends on a variety of factors, including sequence context of the mutation, mutation rate and genotoxic events, such as chemotherapy.
Lower Panel: **CHIP is driven through cell-intrinsic and environmental factors**. Here, the time to acquisition of the CHIP mutation (Mut.) and the change in fitness conferred by the mutation are again determining factors. However, clonal outgrowth is enhanced or enabled by environmental changes (Env. yellow background). Supposed environmental factors are discussed in the main text and include inflammation and other age-related changes.

in *Tet2-/-* HSPCs particularly at sites which gained 5-hydroxymethylcytosine, suggesting TET2 mediated cell-intrinsic changes in HSPCs leading to malignant transformation (Pan et al., 2017).

It is well documented that CHIP mutations can result in inflammation, leading to, for example, enhanced atherosclerosis. In this context, Jaiswal and colleagues (Jaiswal et al., 2014) showed that, when atherosclerosis prone mice were transplanted with *Tet2* knockout bone marrow cells, the development of atherosclerosis was markedly accelerated on the background of a high cholesterol, high-fat diet. In addition, higher levels of pro-inflammatory chemokines could be detected in the serum of these mice. On the molecular level, *Tet2* knockout macrophages, when cultured with low-density lipoprotein, displayed a highly inflammatory transcriptional signature compared to wild type (WT) macrophages, suggesting the involvement of inflammatory signalling in the progression of atherosclerosis on a *Tet2* mutant background. In addition, Fuster and colleagues studied the effects of *Tet2* mutant HSPCs and their progeny in atherosclerosis prone mice deficient in low-density lipoprotein receptor (*Ldlr–/–*) by competitive

transplantation (Fuster et al., 2017). The authors demonstrated that *TET2* loss of function in macrophages exacerbated NLR Family Pyrin Domain Containing 3 (*Nlrp3*) mediated Interleukin 1 beta (IL-1b) production which in turn accelerated atherosclerosis in a context of CHIP, thereby demonstrating that CHIP leads to increased inflammation. Inflammation resulting from mutations in CHIP associated genes is further evidenced by a mouse study, examining the co-operating oncogenic effects of *Jak2V617F* and *Dnmt3a* in HSPCs (Jacquelin et al., 2018). *Dnmt3a* loss on top of the *Jak2V617F* mutation led to the activation of inflammatory signalling, inducing myelofibrosis. A recent study showed that the *Jak2V617F* mutation in HSPCs gave rise to circulating myeloid cells with enhanced pro-inflammatory properties on its own in mouse models of cardiac injury (Sano et al., 2019). These studies demonstrate a role for cell-intrinsic activation of inflammatory pathways as a consequence of CHIP.

Modelling approaches contribute further to evidence for mutations alone being able to explain CHIP. Watson and colleagues (Watson et al., 2020) used PBMC sequencing data from various CHIP studies, analysing 50,000 individuals at varying VAFs. They showed that modelling clone size distributions based on a change of fitness conferred by driver mutations and the probability (time) to acquire these mutations was enough to predict the observed distributions from the collected data sets.

3. Contributing environmental factors towards CHIP

Systemic inflammation from the environment can promote CHIP through, for example, short term inflammatory stress caused by lipopolysaccharides in HSPCs (Cai et al., 2018) (Fig. 1). Murine *Tet2* knockout HSPCs display a survival advantage compared to WT HPSCs during acute inflammation. Following inflammation, *Tet2* knockout HSPCs activated the interleukin 6 (Il6) mediated Stat3/Morbid axis, leading to the upregulation of B-cell lymphoma 2 (*Bcl2*) pro-survival factor and reduced apoptosis in these cells. Therefore, cell-extrinsic factors such as an inflammatory milieu can enhance the competitive fitness of CHIP mutant HSPCs over time.

DNA damage accumulates in aged HSCs and leads to decreased stem cell function. One study linked increased DNA damage directly to exit from the homeostatic quiescent state of HSCs as a response to physiological stresses, explaining the accumulation of DNA damage in aged HSC (Walter et al., 2015). The authors used polyinosinic:polycytidylic acid to mimic viral infections, effectively mounting a type I interferon response (Walter et al., 2015). In addition, a recent study implicated Rad21/Cohesin mediated NFkB signalling in aged HSCs with loss of self-renewal in favour of myeloid biased differentiation in response to inflammatory stimuli (Chen et al., 2019). Inflammation-induced exit from quiescence and ageing-associated inflammation in blood serum and tissue could, therefore, influence the selection of mutant HSPCs carrying CHIP mutations.

A small study in 187 ulcerative colitis (UC) patients, an autoimmune disease characterised by increased levels of proinflammatory cytokines with an average onset age before 30 years, analysed targeted PBMC sequencing for CHIP mutations (Zhang et al., 2019). Albeit patient numbers being small and the lack of a validation cohort, the study revealed *DNMT3A* and *PPM1D* as the most prevalent mutations with a lower incidence of *TET2* mutations. In this cohort, overall CHIP was slightly higher in UC patients, with *DNMT3A* mutant patients revealing significantly higher levels of serum interferon-gamma (IFNg), but not tumour necrosis factor-alpha. Interestingly, *DNMT3A* VAF was a significant contributor to increased IFNg levels, suggesting that increased IFNg might select for *DNMT3A* mutations in UC. Given that the onset of UC mostly occurs before the age of 30 years - compared to the onset of CHIP occurring at least two decades later, this study might provide evidence of a pro-inflammatory milieu playing a role in CHIP onset (Ha et al., 2010). However, in a minority of UC patients, disease occurs at the age of 50 years or older. In this context, inflammation might occur first, resulting from, for example, *DNMT3A* mutant T-cell clones (Thomas et al., 2010) or other CHIP related inflammatory processes as discussed above.

4. Associations with CHIP

Most CHIP studies describe associations with age-related multimorbidities or ageing factors, leaving uncertainty over cell-intrinsic versus environmental factors and their contributions to CHIP.

A key determinant of CHIP is the presence of recurrent driver mutations that are functionally well described; however, many samples frequently present with not known causal variants - a phenomenon known as clonal haematopoiesis with unknown drivers (CH-UD) (Genovese et al., 2014). Genovese and colleagues performed whole-exome sequencing of PBMCs on 11,845 participants, where the majority had not known putative driver mutations, with 1333 participants displaying 1 mutation, 313 participants harbouring 2 mutations, and 272 with 3 to 18 somatic mutations in total (Genovese et al., 2014). The authors then defined CH-UD based on the mutational burden in passenger genes alone, rather than on the identity of the mutations. In some participants without known driver mutations, further analysis and deeper sequencing eventually revealed a candidate variant (Genovese et al., 2014), suggesting that detection sensitivity might have been limiting in the first instance. The absence of canonical CHIP variants might also be explained by copy-number alterations of affected genes; however, it is unlikely that these factors explain all cases. Indeed, some have hypothesised CH-UD may be linked to reduced HSC fitness with age which results in increased oligoclonality through a depletion of the HSC pool (Gibson and Steensma, 2018).

Robertson and colleagues (Robertson et al., 2019) recently showed an increase in epigenetic age, a correlate of biological age, in CHIP carriers when compared to individuals without detectable CHIP. In this study, CHIP mutations were annotated in the Lothian Birth Cohorts (LBCs) of 1921 (n = 550) and 1936 (n = 1091), two independent, longitudinal studies in the elderly using whole-genome sequencing. Epigenetic clock analysis was then performed on 450 K methylation arrays. Increased epigenetic age was noted when considering all CHIP mutations together, and *TET2* and *DNMT3A* mutations individually. These effects were larger than the known sex differences in age acceleration (male > female) in either cohort. Moreover, VAF was positively correlated with accelerated epigenetic age, suggesting a link to clone size, which could be driven by intrinsic or environmental factors.

Zink and colleagues performed deep whole-genome sequencing (WGS) in 11,262 Icelanders and identified 1403 cases of CHIP at 2% VAF, irrespective of driver mutation status (Zink et al., 2017). Overall CHIP in this cohort was much more common compared to other studies, with 50% of participants over the age of 85 being carriers, showing similar somatic mutation patterns as previously reported (*TET2*, *DNMT3A*, *ASXL1*, *PPM1D*). In this cohort, CHIP mutations were associated with reduced telomere length (Zink et al., 2017). This finding complements the notion that epigenetic age is altered in CHIP, suggesting that proliferation might be a feature contributing to cell-intrinsic ageing factors and systemic ageing. Interestingly, known driver mutations were only apparent in a fraction of CHIP carriers. Using modelling approaches, the authors suggested that some clones have arisen in the absence of mutations as a result of neutral drift, which would only act on a small number of active HSCs. However, the majority of CHIP cases in the absence of mutations remained unexplained, suggesting environmental influences.

Further evidence for clonal outgrowth due to increased age comes from a study where WGS of PBMCs from a 115-year-old woman was performed (Holstege et al., 2014). The authors detected 450 somatic mutations, which were reported as passenger mutations, leading to oligoclonal haematopoiesis (Holstege et al., 2014). The authors suggested that the finite lifespan of HSCs leads to CHIP rather than the acquisition of driver mutations. Whether HSC lifespan is mainly

regulated by environmental (e.g. cytokines promoting proliferation) or cell-intrinsic factors (e.g. telomere length) remains to be seen. During ageing, senescence due to telomere shortening or other cues has been described as a main driver of a proinflammatory environment (Acosta et al., 2008; Coppé et al., 2008; Kuilman et al., 2008). Telomere shortening in bone marrow stromal cells was correlated with a dysfunctional haematopoietic environment and increased cytokines in ageing context (Ju et al., 2007). Whether senescence and SASP play a significant role during ageing of the blood compartment remains to be conclusively elucidated.

5. Concluding remarks

It is becoming increasingly clear that certain mutations lead to CHIP. For *TET2* and *SRSFP95H* mutations, myeloid bias was associated with CHIP or the initiation of myelodysplastic/myeloproliferative syndrome, whereas for *DNMT3A* multipotent stem cell origin was described in the context of CHIP (Buscarlet et al., 2018; Smeets et al., 2018). However, competition or cooperation between clones of distinct sizes, lineage origin and different types of mutations is currently unknown. For example, could an early clone create an environment favouring outgrowth of certain other clones over time, by, for example, promoting inflammation?

Although evidence is increasing for mutations driving CHIP, several studies also suggest clonal outgrowth can appear in the absence of driver mutations. One such scenario could be due to a selection based on transcriptional phenotypes in the absence of CHIP, as suggested for ageing mouse HSCs. In one study, only a minority subpopulation of HSCs developed transcriptional signatures commonly associated with HSC ageing such as *Tp53* (Kirschner et al., 2017). The majority of HSCs showed a transcriptome similar to that of young HSCs, suggesting heterogeneous ageing phenotypes on the transcriptional level (Kirschner et al., 2017). Whether proliferation of HSC subpopulations was driven cell-intrinsically, or whether HSC subpopulations were simply being kept in a low proliferative state over time, gaining an advantage over exhausted, pro-proliferative HSC populations with increased age, remains to be elucidated.

Author contribution

K.K. and T.C. conceived the review. M.T.T., N.A.R., K.K. and T.C. wrote the manuscript.

Declaration of Competing Interest

The authors declare no competing interests

Acknowledgements

K.K. is funded by a John Goldman Fellowship sponsored by Leukaemia U.K. (2019/JGF/003). M.T.T. and N.A.R. are supported by MRC funded Ph.D. studentships (MR/N013166/1). T.C. is supported by Chancellor's Fellowship held at the University of Edinburgh.

References

Acosta, J.C., O'Loghlen, A., Banito, A., Guijarro, M.V., Augert, A., Raguz, S., Fumagalli, M., Da Costa, M., Brown, C., Popov, N., Takatsu, Y., Melamed, J., d'Adda di Fagagna, F., Bernard, D., Hernando, E., Gil, J., 2008. Chemokine signaling via the CXCR2 receptor reinforces senescence. Cell 133, 1006–1018. https://doi.org/10.1016/j.cell.2008.03.038.

Bowman, R.L., Busque, L., Levine, R.L., 2018. Clonal hematopoiesis and evolution to hematopoietic malignancies. Cell Stem Cell 22, 157–170. https://doi.org/10.1016/j.stem.2018.01.011.

Buscarlet, M., Provost, S., Zada, Y.F., Bourgoin, V., Mollica, L., Dubé, M.-P., Busque, L., 2018. Lineage restriction analyses in CHIP indicate myeloid bias for TET2 and multipotent stem cell origin for DNMT3A. Blood 132, 277–280. https://doi.org/10.1182/blood-2018-01-829937.

Cai, Z., Kotzin, J.J., Ramdas, B., Chen, S., Nelanuthala, S., Palam, L.R., Pandey, R., Mali, R.S., Liu, Y., Kelley, M.R., Sandusky, G., Mohseni, M., Williams, A., Henao-Mejia, J., Kapur, R., 2018. Inhibition of inflammatory signaling in Tet2 mutant preleukemic cells mitigates stress-induced abnormalities and clonal hematopoiesis. Cell Stem Cell 23, 833–849. https://doi.org/10.1016/j.stem.2018.10.013. e5.

Challen, G.A., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J.S., Bock, C., Vasanthakumar, A., Gu, H., Xi, Y., Liang, S., Lu, Y., Darlington, G.J., Meissner, A., Issa, J.-P.J., Godley, L.A., Li, W., Goodell, M.A., 2011. Dnmt3a is essential for hematopoietic stem cell differentiation. Nat. Genet. 44, 23–31. https://doi.org/10.1038/ng.1009.

Chandra, T., Kirschner, K., 2016. Chromosome organisation during ageing and senescence. Curr. Opin. Cell Biol. 40, 161–167. https://doi.org/10.1016/j.ceb.2016.03.020.

Chen, E., Staudt, L.M., Green, A.R., 2012. Janus kinase deregulation in leukemia and lymphoma. Immunity 36, 529–541. https://doi.org/10.1016/j.immuni.2012.03.017.

Chen, Z., Amro, E.M., Becker, F., Hölzer, M., Rasa, S.M.M., Njeru, S.N., Han, B., Di Sanzo, S., Chen, Y., Tang, D., Tao, S., Haenold, R., Groth, M., Romanov, V.S., Kirkpatrick, J.M., Kraus, J.M., Kestler, H.A., Marz, M., Ori, A., Neri, F., Rudolph, K.L., 2019. Cohesin-mediated NF-κB signaling limits hematopoietic stem cell self-renewal in aging and inflammation. J. Exp. Med. 216, 152–175. https://doi.org/10.1084/jem.20181505.

Cho, R.H., Sieburg, H.B., Muller-Sieburg, C.E., 2008. A new mechanism for the aging of hematopoietic stem cells: aging changes the clonal composition of the stem cell compartment but not individual stem cells. Blood 111, 5553–5561. https://doi.org/10.1182/blood-2007-11-123547.

Coombs, C.C., Zehir, A., Devlin, S.M., Kishtagari, A., Syed, A., Jonsson, P., Hyman, D.M., Solit, D.B., Robson, M.E., Baselga, J., Arcila, M.E., Ladanyi, M., Tallman, M.S., Levine, R.L., Berger, M.F., 2017. Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. Cell Stem Cell 21, 374–382. https://doi.org/10.1016/j.stem.2017.07.010. e4.

Coppé, J.-P., Patil, C.K., Rodier, F., Sun, Y., Muñoz, D.P., Goldstein, J., Nelson, P.S., Desprez, P.-Y., Campisi, J., 2008. Senescence-associated secretory phenotypes reveal cell-nonautonomous functions of oncogenic RAS and the p53 tumor suppressor. PLoS Biol. 6, 2853–2868. https://doi.org/10.1371/journal.pbio.0060301.

Cypris, O., Božić, T., Wagner, W., 2019. Chicken or egg: is clonal hematopoiesis primarily caused by genetic or epigenetic aberrations? Front. Genet. 10, 785. https://doi.org/10.3389/fgene.2019.00785.

de Magalhães, J.P., 2013. How ageing processes influence cancer. Nat. Rev. Cancer 13, 357–365. https://doi.org/10.1038/nrc3497.

Dorsheimer, L., Assmus, B., Rasper, T., Ortmann, C.A., Ecke, A., Abou-El-Ardat, K., Schmid, T., Brüne, B., Wagner, S., Serve, H., Hoffmann, J., Seeger, F., Dimmeler, S., Zeiher, A.M., Rieger, M.A., 2019. Association of mutations contributing to clonal hematopoiesis with prognosis in chronic ischemic heart failure. JAMA Cardiol. 4, 25–33. https://doi.org/10.1001/jamacardio.2018.3965.

Fuster, J.J., MacLauchlan, S., Zuriaga, M.A., Polackal, M.N., Ostriker, A.C., Chakraborty, R., Wu, C.-L., Sano, S., Muralidharan, S., Rius, C., Vuong, J., Jacob, S., Muralidhar, V., Robertson, A.A.B., Cooper, M.A., Andrés, V., Hirschi, K.K., Martin, K.A., Walsh, K., 2017. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. Science 355, 842–847. https://doi.org/10.1126/science.aag1381.

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., Purcell, S.M., Svantesson, O., Landén, M., Höglund, M., Lehmann, S., Gabriel, S.B., Moran, J.L., Lander, E.S., Sullivan, P.F., Sklar, P., McCarroll, S.A., 2014. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. 371, 2477–2487. https://doi.org/10.1056/NEJMoa1409405.

Gibson, C.J., Steensma, D.P., 2018. New insights from studies of clonal hematopoiesis. Clin. Cancer Res. 24, 4633–4642. https://doi.org/10.1158/1078-0432.CCR-17-3044.

Ha, C.Y., Newberry, R.D., Stone, C.D., Ciorba, M.A., 2010. Patients with late-adult-onset ulcerative colitis have better outcomes than those with early onset disease. Clin. Gastroenterol. Hepatol. 8, 682–687. https://doi.org/10.1016/j.cgh.2010.03.022. e1.

Holstege, H., Pfeiffer, W., Sie, D., Hulsman, M., Nicholas, T.J., Lee, C.C., Ross, T., Lin, J., Miller, M.A., Ylstra, B., Meijers-Heijboer, H., Brugman, M.H., Staal, F.J.T., Holstege, G., Reinders, M.J.T., Harkins, T.T., Levy, S., Sistermans, E.A., 2014. Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. Genome Res. 24, 733–742. https://doi.org/10.1101/gr.162131.113.

Hsu, J.I., Dayaram, T., Tovy, A., De Braekeleer, E., Jeong, M., Wang, F., Zhang, J., Heffernan, T.P., Gera, S., Kovacs, J.J., Marszalek, J.R., Bristow, C., Yan, Y., Garcia-Manero, G., Kantarjian, H., Vassiliou, G., Futreal, P.A., Donehower, L.A., Takahashi, K., Goodell, M.A., 2018. PPM1D mutations drive clonal hematopoiesis in response to cytotoxic chemotherapy. Cell Stem Cell 23, 700–713. https://doi.org/10.1016/j.stem.2018.10.004. e6.

Inoue, D., Bradley, R.K., Abdel-Wahab, O., 2016. Spliceosomal gene mutations in myelodysplasia: molecular links to clonal abnormalities of hematopoiesis. Genes Dev. 30, 989–1001. https://doi.org/10.1101/gad.278424.116.

Jacquelin, S., Straube, J., Cooper, L., Vu, T., Song, A., Bywater, M., Baxter, E., Heidecker, M., Wackrow, B., Porter, A., Ling, V., Green, J., Austin, R., Kazakoff, S., Waddell, N., Hesson, L.B., Pimanda, J.E., Stegelmann, F., Bullinger, L., Döhner, K., Lane, S.W., 2018. Jak2V617F and Dnmt3a loss cooperate to induce myelofibrosis through activated enhancer-driven inflammation. Blood 132, 2707–2721. https://doi.org/10.1182/blood-2018-04-846220.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., Higgins, J.M., Moltchanov, V., Kuo, F.C., Kluk, M.J., Henderson, B., Kinnunen, L., Koistinen, H.A., Ladenvall, C., Getz, G., Correa, A., Ebert, B.L., 2014. Age-related clonal hematopoiesis associated with

adverse outcomes. N. Engl. J. Med. 371, 2488–2498. https://doi.org/10.1056/NEJMoa1408617.

Jaiswal, S., Natarajan, P., Silver, A.J., Gibson, C.J., Bick, A.G., Shvartz, E., McConkey, M., Gupta, N., Gabriel, S., Ardissino, D., Baber, U., Mehran, R., Fuster, V., Danesh, J., Frossard, P., Saleheen, D., Melander, O., Sukhova, G.K., Neuberg, D., Libby, P., Ebert, B.L., 2017. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. N. Engl. J. Med. 377, 111–121. https://doi.org/10.1056/NEJMoa1701719.

Jeong, M., Park, H.J., Celik, H., Ostrander, E.L., Reyes, J.M., Guzman, A., Rodriguez, B., Lei, Y., Lee, Y., Ding, L., Guryanova, O.A., Li, W., Goodell, M.A., Challen, G.A., 2018. Loss of dnmt3a immortalizes hematopoietic stem cells in vivo. Cell Rep. 23, 1–10. https://doi.org/10.1016/j.celrep.2018.03.025.

Ju, Z., Jiang, H., Jaworski, M., Rathinam, C., Gompf, A., Klein, C., Trumpp, A., Rudolph, K.L., 2007. Telomere dysfunction induces environmental alterations limiting hematopoietic stem cell function and engraftment. Nat. Med. 13, 742–747. https://doi.org/10.1038/nm1578.

Kahlem, P., Dörken, B., Schmitt, C.A., 2004. Cellular senescence in cancer treatment: friend or foe? J. Clin. Invest. 113, 169–174. https://doi.org/10.1172/JCI20784.

Kahn, J.D., Miller, P.G., Silver, A.J., Sellar, R.S., Bhatt, S., Gibson, C., McConkey, M., Adams, D., Mar, B., Mertins, P., Fereshetian, S., Krug, K., Zhu, H., Letai, A., Carr, S.A., Doench, J., Jaiswal, S., Ebert, B.L., 2018. PPM1D-truncating mutations confer resistance to chemotherapy and sensitivity to PPM1D inhibition in hematopoietic cells. Blood 132, 1095–1105. https://doi.org/10.1182/blood-2018-05-850339.

Kirschner, K., Chandra, T., Kiselev, V., Flores-Santa Cruz, D., Macaulay, I.C., Park, H.J., Li, J., Kent, D.G., Kumar, R., Pask, D.C., Hamilton, T.L., Hemberg, M., Reik, W., Green, A.R., 2017. Proliferation drives aging-related functional decline in a subpopulation of the hematopoietic stem cell compartment. Cell Rep. 19, 1503–1511. https://doi.org/10.1016/j.celrep.2017.04.074.

Kirschner, K., Samarajiwa, S.A., Cairns, J.M., Menon, S., Pérez-Mancera, P.A., Tomimatsu, K., Bermejo-Rodriguez, C., Ito, Y., Chandra, T., Narita, Masako, Lyons, S.K., Lynch, A.G., Kimura, H., Ohbayashi, T., Tavaré, S., Narita, Masashi, 2015. Phenotype specific analyses reveal distinct regulatory mechanism for chronically activated p53. PLoS Genet. 11, e1005053. https://doi.org/10.1371/journal.pgen.1005053.

Ko, M., Bandukwala, H.S., An, J., Lamperti, E.D., Thompson, E.C., Hastie, R., Tsangaratou, A., Rajewsky, K., Koralov, S.B., Rao, A., 2011. Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. Proc. Natl. Acad. Sci. U.S.A. 108, 14566–14571. https://doi.org/10.1073/pnas.1112317108.

Kuilman, T., Michaloglou, C., Vredeveld, L.C.W., Douma, S., van Doorn, R., Desmet, C.J., Aarden, L.A., Mooi, W.J., Peeper, D.S., 2008. Oncogene-induced senescence relayed by an interleukin-dependent inflammatory network. Cell 133, 1019–1031. https://doi.org/10.1016/j.cell.2008.03.039.

Li, Z., Cai, X., Cai, C.-L., Wang, J., Zhang, W., Petersen, B.E., Yang, F.-C., Xu, M., 2011. Deletion of Tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. Blood 118, 4509–4518. https://doi.org/10.1182/blood-2010-12-325241.

Lu, R., Wang, P., Parton, T., Zhou, Y., Chrysovergis, K., Rockowitz, S., Chen, W.-Y., Abdel-Wahab, O., Wade, P.A., Zheng, D., Wang, G.G., 2016. Epigenetic perturbations by Arg882-Mutated DNMT3A potentiate aberrant stem cell gene-expression program and acute leukemia development. Cancer Cell 30, 92–107. https://doi.org/10.1016/j.ccell.2016.05.008.

Mayle, A., Yang, L., Rodriguez, B., Zhou, T., Chang, E., Curry, C.V., Challen, G.A., Li, W., Wheeler, D., Rebel, V.I., Goodell, M.A., 2015. Dnmt3a loss predisposes murine hematopoietic stem cells to malignant transformation. Blood 125, 629–638. https://doi.org/10.1182/blood-2014-08-594648.

McKerrell, T., Park, N., Moreno, T., Grove, C.S., Ponstingl, H., Stephens, J., Understanding Society Scientific Group, Crawley, C., Craig, J., Scott, M.A., Hodkinson, C., Baxter, J., Rad, R., Forsyth, D.R., Quail, M.A., Zeggini, E., Ouwehand, W., Varela, I., Vassiliou, G.S., 2015. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. Cell Rep. 10, 1239–1245. https://doi.org/10.1016/j.celrep.2015.02.005.

Moran-Crusio, K., Reavie, L., Shih, A., Abdel-Wahab, O., Ndiaye-Lobry, D., Lobry, C., Figueroa, M.E., Vasanthakumar, A., Patel, J., Zhao, X., Perna, F., Pandey, S., Madzo, J., Song, C., Dai, Q., He, C., Ibrahim, S., Beran, M., Zavadil, J., Nimer, S.D., Levine, R.L., 2011. Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. Cancer Cell 20, 11–24. https://doi.org/10.1016/j.ccr.2011.06.001.

Pan, F., Wingo, T.S., Zhao, Z., Gao, R., Makishima, H., Qu, G., Lin, L., Yu, M., Ortega, J.R., Wang, Jiapeng, Nazha, A., Chen, L., Yao, B., Liu, C., Chen, S., Weeks, O., Ni, H., Phillips, B.L., Huang, S., Wang, Jianlong, Xu, M., 2017. Tet2 loss leads to hypermutagenicity in haematopoietic stem/progenitor cells. Nat. Commun. 8, 15102. https://doi.org/10.1038/ncomms15102.

Quivoron, C., Couronné, L., Della Valle, V., Lopez, C.K., Plo, I., Wagner-Ballon, O., Do Cruzeiro, M., Delhommeau, F., Arnulf, B., Stern, M.-H., Godley, L., Opolon, P., Tilly, H., Solary, E., Duffourd, Y., Dessen, P., Merle-Beral, H., Nguyen-Khac, F., Fontenay, M., Vainchenker, W., Bernard, O.A., 2011. TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer Cell 20, 25–38. https://doi.org/10.1016/j.ccr.2011.06.003.

Robertson, N.A., Hillary, R.F., McCartney, D.L., Terradas-Terradas, M., Higham, J., Sproul, D., Deary, I.J., Kirschner, K., Marioni, R.E., Chandra, T., 2019. Age-related clonal haematopoiesis is associated with increased epigenetic age. Curr. Biol. 29, R786–R787. https://doi.org/10.1016/j.cub.2019.07.011.

Sano, S., Wang, Y., Yura, Y., Sano, M., Oshima, K., Yang, Y., Katanasaka, Y., Min, K.-D., Matsuura, S., Ravid, K., Mohi, G., Walsh, K., 2019. JAK2V617F -Mediated clonal hematopoiesis accelerates pathological remodeling in murine heart failure. JACC Basic Transl. Sci. 4, 684–697. https://doi.org/10.1016/j.jacbts.2019.05.013.

Smeets, M.F., Tan, S.Y., Xu, J.J., Anande, G., Unnikrishnan, A., Chalk, A.M., Taylor, S.R., Pimanda, J.E., Wall, M., Purton, L.E., Walkley, C.R., 2018. Srsf2P95H initiates myeloid bias and myelodysplastic/myeloproliferative syndrome from hemopoietic stem cells. Blood 132, 608–621. https://doi.org/10.1182/blood-2018-04-845602.

Thomas, R., Gamper, C., Powell, J., Wells, A., 2010. DNMT3a mediates lineage-specific, de novo DNA methylation at the ifng promoter and contributes to ifng gene silencing in Th2, Th17 and Treg cells (88.15). J. Immunol. 184 88.15-88.15.

Walter, D., Lier, A., Geiselhart, A., Thalheimer, F.B., Huntscha, S., Sobotta, M.C., Moehrle, B., Brocks, D., Bayindir, I., Kaschutnig, P., Muedder, K., Klein, C., Jauch, A., Schroeder, T., Geiger, H., Dick, T.P., Holland-Letz, T., Schmezer, P., Lane, S.W., Rieger, M.A., Milsom, M.D., 2015. Exit from dormancy provokes DNA-damage-induced attrition in haematopoietic stem cells. Nature 520, 549–552. https://doi.org/10.1038/nature14131.

Wang, J., Li, Z., He, Y., Pan, F., Chen, S., Rhodes, S., Nguyen, L., Yuan, J., Jiang, L., Yang, X., Weeks, O., Liu, Z., Zhou, J., Ni, H., Cai, C.-L., Xu, M., Yang, F.-C., 2014. Loss of Asxl1 leads to myelodysplastic syndrome-like disease in mice. Blood 123, 541–553. https://doi.org/10.1182/blood-2013-05-500272.

Watson, C.J., Papula, A.L., Poon, G.Y.P., Wong, W.H., Young, A.L., Druley, T.E., Fisher, D.S., Blundell, J.R., 2020. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science 367, 1449–1454. https://doi.org/10.1126/science.aay9333.

Zhang, C.R.C., Nix, D., Gregory, M., Ciorba, M.A., Ostrander, E.L., Newberry, R.D., Spencer, D.H., Challen, G.A., 2019. Inflammatory cytokines promote clonal hematopoiesis with specific mutations in ulcerative colitis patients. Exp. Hematol. 80, 36–41. https://doi.org/10.1016/j.exphem.2019.11.008. e3.

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., Jonasson, J.G., Tryggvadottir, L., Jonsson, T., Helgason, A., Gylfason, A., Sulem, P., Rafnar, T., Thorsteinsdottir, U., Gudbjartsson, D.F., Masson, G., Stefansson, K., 2017. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. Blood 130, 742–752. https://doi.org/10.1182/blood-2017-02-769869.

Appendix 7: Manuscript – Age Related Clonal Haematopoiesis is Associated with Increased Epigenetic Age (2019)

Correspondence

Age-related clonal haemopoiesis is associated with increased epigenetic age

Neil A. Robertson[1], Robert F. Hillary[2], Daniel L. McCartney[2], Maria Terradas-Terradas[3], Jonathan Higham[1], Duncan Sproul[1,4], Ian J. Deary[5,6,*], Kristina Kirschner[3,*], Riccardo E. Marioni[2,5,*], and Tamir Chandra[1,*]

Age-related clonal haemopoiesis (ARCH) in healthy individuals was initially observed through an increased skewing in X-chromosome inactivation [1]. More recently, several groups reported that ARCH is driven by somatic mutations [2], with the most prevalent ARCH mutations being in the *DNMT3A* and *TET2* genes, previously described as drivers of myeloid malignancies. ARCH is associated with an increased risk for haematological cancers [2]. ARCH also confers an increased risk for non-haematological diseases, such as cardiovascular disease, atherosclerosis, and chronic ischemic heart failure, for which age is a main risk factor [3,4]. Whether ARCH is linked to accelerated ageing has remained unexplored. The most accurate and commonly used tools to measure age acceleration are epigenetic clocks: they are based on age-related methylation differences at specific CpG sites [5]. Deviations from chronological age towards an increased epigenetic age have been associated with increased risk of earlier mortality and age-related morbidities [5,6]. Here we present evidence of accelerated epigenetic age in individuals with ARCH.

The Lothian Birth Cohorts (LBCs) of 1921 and 1936 are two longitudinal studies of ageing [7]. Participants have been followed up every ~3 years, each for five waves, from the age of 70 (LBC1936) and 79 (LBC1921). Participants were community dwelling, relatively healthy, and mostly lived in the City of Edinburgh or its surrounding area when recruited.

Whole-blood DNA methylation levels were assessed using the Illumina HumanMethylation450 BeadChip (Supplemental Experimental Procedures). Genomic variants were determined in 1,136 LBC participants (n = 870 from wave 1 at mean age 70 years in LBC1936; n = 101 and n = 165 at mean ages 79 and 87, respectively, in LBC1921) with whole-genome sequencing (WGS) and methylation data. WGS data were aligned with Burrows-Wheeler Aligner and processed for duplicate mapping reads with samblaster (genome coverage of 34.3 reads). Single-nucleotide variants and short indels were called with MuTect (v3.8) before annotation using the Ensembl Variant Effect Predictor alongside the Cosmic database of coding mutations (v86). ARCH variants were classified as per Jaiswal *et al*. [2].

Epigenetic age acceleration was calculated online (https://dnamage.genetics.ucla.edu/home). We considered the Intrinsic Epigenetic Age Acceleration (IEAA, hereafter referred to as Horvath age acceleration) measure, which is an adapted version of the original Horvath clock that controls for white blood cell proportions [6]. Epigenetic age estimates were regressed on chronological age to yield age acceleration residuals. Linear regression adjusting for sex, imputed white blood cell proportions (monocytes, natural killer (NK), $CD4^+$ T, $CD8^+$ T, and B cells), and methylation processing batch was used to determine the association between ARCH status and Age Acceleration. All analyses were conducted in R v3.5.0.

Of the ten most prevalent ARCH mutations [2], we had sufficient sample size and sequencing depth to annotate the top six in the LBCs. We identified 73 participants (from 1,136) with ARCH (6%; Figure 1A). The gene-specific prevalence ranged between 1 and 36 cases with ARCH-variant allele frequencies ranging from 0.034 to 0.677 (Figure 1B). Mutations in *TET2* were exclusively frameshift and mutations detected in *JAK2* (all V617F), *SF3B1* and *TP53* were exclusively missense. ARCH status was associated with a significant increase in Horvath age acceleration: the increase was 4.5 (SE 0.9) years in LBC1936, and 3.7 (SE 1.2) years in LBC1921 ($p = 2.3 \times 10^{-6}$ and 2.5×10^{-3}, respectively; Figure 1C and Table S1). Compared with non-ARCH carriers, those with *TET2* mutations had a 6.1 (SE 2.2) year and 6.4 (SE 1.9) year increase in Horvath age acceleration in LBC1936 and LBC1921 (p = 0.004 and p = 0.001), respectively. Those with *DNMT3A* mutations had 3.8 (SE 1.2) years increase in LBC1936, and 3.0 (SE 1.9) years in LBC1921 (p = 0.002 and p = 0.11), respectively (Figure 1D). These effect sizes are much larger than the sex differences in Horvath age acceleration, which were 1.8 (SE 0.4) years for men in LBC1936 ($p = 5.1 \times 10^{-5}$), and 1.0 years (SE 0.8) in LBC1921 (p = 0.18) (Figure 1D and Table S1).

We also considered age acceleration estimates from four additional epigenetic clocks: Extrinsic (Hannum) Epigenetic Age (EEAA) [6], PhenoAge [8], GrimAge [9] and Zhang Age [10] (Figure 1E,F and Figure S1A–F). Briefly, ARCH status was linked to increased EEAA, PhenoAge, GrimAge and ZhangAge, acceleration in LBC1921 (effect sizes: 1.9 years, 3.7 years, 2.8 years and 0.8 years with p = 0.16, 0.014, 9.6×10^{-4}, and 3.5×10^{-3}, respectively). In LBC1936 there was a modest association between ARCH and increased EEAA and ZhangAge (2.3 years and 0.5 years, p = 0.012 and 4.4×10^{-3}) but no association with PhenoAge or GrimAge acceleration (p = 0.32 and 0.99, respectively). There was no consistent association between ARCH status and white cell count proportions across the two cohorts: a lower proportion of NK cells was linked with ARCH carrier status in LBC1936 (odds ratio per SD of cell counts, 0.57 95% CI [0.37, 0.84]), while a higher B cell proportion was associated with ARCH status in LBC1921 (OR 1.37 [1.01, 1.94]).

We observed associations between ARCH and epigenetic age acceleration in the independent LBCs of 1921 and 1936, where the WGS data and the DNA methylation data were processed together using identical protocols. Although we examined multiple epigenetic clocks in relation to ARCH status, it is possible that the effect sizes may vary by the quality control approach applied to the methylation data. Additional replication from other cohorts would further strengthen the magnitude and generalisability of the associations. Our results could indicate ARCH as an underlying cause for systemic ageing, explaining its link to non-haematological, age-related diseases.

SUPPLEMENTAL INFORMATION

Supplemental Information includes one figure, one table and experimental procedures and can be found with this article online at https://doi.org/10.1016/j.cub.2019.07.011.

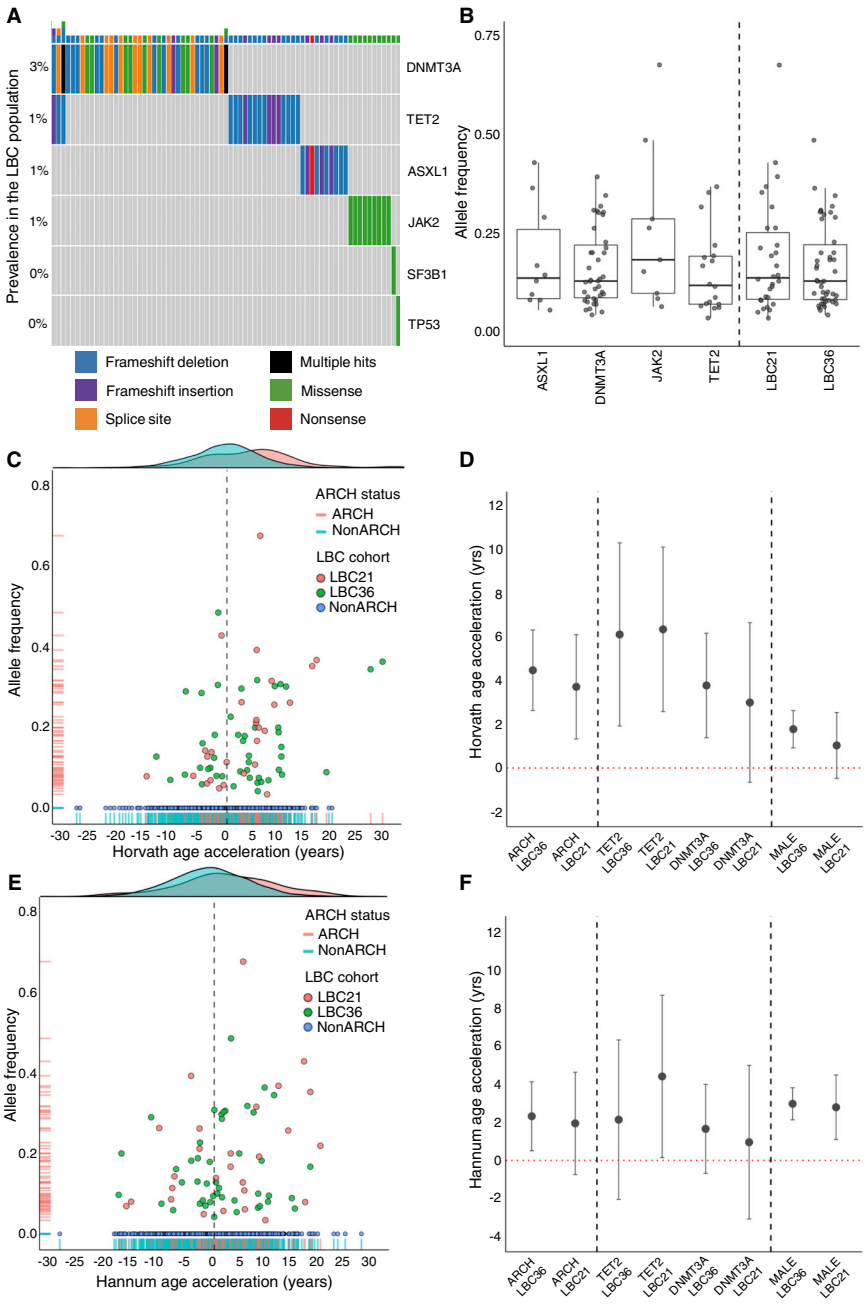


Figure 1. ARCH variants discovered in Lothian Birth Cohort (LBC) participants and their effects on epigenetic age as shown in the Horvath (IEAA) and Hannum (EEAA) clocks.
(A) Oncoplot showing variant types within the ARCH positive subset of the LBC. This subset represents 73 participants (6% of 1,136 total) where one or more described somatic variants were detected in the six most prevalent ARCH-associated genes. (B) Box plot describing the distribution of allele frequencies in all detected somatic ARCH variants. Genes with a single variant not shown are *TP53* and *SF3B1* (allele frequencies of 0.089 and 0.257, respectively). The overall distribution of allele frequencies by LBC cohort (LBC1921/LBC1936) is also displayed. (C) Scatter plot of Horvath age acceleration (IEAA; years) for individual LBC participants against the allele frequency of their somatic ARCH variant in both LBC1921 (orange dots, net 3.7 years; $p = 2.5 \times 10^{-3}$) and LBC1936 (green dots, net 4.5 years; $p = 2.3 \times 10^{-6}$) cohorts. Density plot highlighting the shift in distribution of Horvath age acceleration between ARCH-positive (orange) and -negative participant (turquoise) groups. Non-ARCH carriers (blue dots). (D) Plot showing net IEAA in ARCH (with 95% confidence intervals). The effect of sex (male versus female) on epigenetic ageing within the LBC is shown for comparison. (E) Scatter plot showing the Hannum age acceleration (EEAA; years) against the allele frequency of ARCH variants in both LBC1921 (orange dots, net 1.9 years; $p = 0.16$) and LBC1936 (green dots, net 2.3 years; $p = 0.01$) cohorts. Density plot highlighting shift in distribution of EEAA between ARCH-positive (orange) and -negative participant (turquoise) groups. Non-ARCH carriers (blue dots). (F) Plot showing the net EEAA in ARCH (with 95% confidence intervals). The effect of sex (male versus female) on epigenetic ageing within the LBC is shown for comparison.

REFERENCES

1. Busque, L., Mio, R., Mattioli, J., Brais, E., Blais, N., Lalonde, Y., Maragh, M., and Gilliland, D.G. (1996). Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. Blood *88*, 59–65.
2. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., *et al*. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.
3. Dorsheimer, L., Assmus, B., Rasper, T., Ortmann, C.A., Ecke, A., Abou-El-Ardat, K., Schmid, T., Brüne, B., Wagner, S., Serve, H., *et al*. (2019). Association of mutations contributing to clonal hematopoiesis with prognosis in chronic ischemic heart failure. JAMA Cardiol. *4*, 25–33.
4. Jaiswal, S., Natarajan, P., Silver, A.J., Gibson, C.J., Bick, A.G., Shvartz, E., McConkey, M., Gupta, N., Gabriel, S., Ardissino, D., *et al*. (2017). Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. N. Engl. J. Med. *377*, 111–121.
5. Horvath, S., and Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat. Rev. Genet. *19*, 371–384.
6. Chen, B.H., Marioni, R.E., Colicino, E., Peters, M.J., Ward-Caviness, C.K., Tsai, P.-C., Roetker, N.S., Just, A.C., Demerath, E.W., Guan, W., *et al*. (2016). DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging *8*, 1844–1865.
7. 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.
8. Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D., Li, Y., *et al*. (2018). An epigenetic biomarker of aging for lifespan and healthspan. Aging *10*, 573–591.
9. Lu, A.T, Quach, A., Wilson, J.G., Reiner, A.P., Aviv, A., Raj, K., Hou, L., Baccarelli, A.A., Li, Y., Stewart, J.D., *et al*. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging *11*, 303–327.
10. Zhang, Q., Vallerga, C., Walker, R., Lin, T, Henders, A., Montgomery, G., He, J., Fan, D., Fowdar, J., Kennedy, M., *et al*. (2018). Improved prediction of chronological age from DNA methylation limits it as a biomarker of ageing. bioRxiv https://doi.org/10.1101/327890.

[1]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK. [2]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK. [3]Institute of Cancer Sciences, University of Glasgow, Glasgow, G61 1BD, UK. [4]Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, UK. [5]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, EH8 9JZ, UK. [6]Department of Psychology, University of Edinburgh, Edinburgh, EH8 9JZ, UK.
*E-mail: (I.J.D.), (K.K.), (R.E.M.), (T.C.)