Clonal dynamics of haematopoiesis across the human lifespan

Dr Emily Louise Mitchell

Magdalene College

University of Cambridge

March 2022

Thesis submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the Biological Sciences Degree Committee of 60,000 words.

Summary

Clonal dynamics of haematopoiesis across the human lifespan

Emily Mitchell

The haematopoietic system manifests several age-associated phenotypes including anaemia; loss of regenerative capacity, especially in the face of insults such as infection, chemotherapy or blood loss; and increased risk of clonal haematopoiesis and blood cancers. The cellular alterations that underpin these age-related phenotypes, which typically manifest in individuals aged over 70, remain elusive. In my thesis I have aimed to investigate whether changes in HSC population structure with age might underlie any aspects of haematopoietic system ageing. In addition, I have investigated the impact of chemotherapeutic perturbations on haematopoietic stem cell mutation burden and clonal dynamics.

To answer the ageing question, I have sequenced 3579 genomes from single-cell-derived colonies of haematopoietic stem cell/multipotent progenitors (HSC/MPPs) from 10 haematologically normal subjects aged 0-81 years. HSC/MPPs accumulated 17 somatic mutations/year after birth with no increased rate of mutation accumulation in the elderly. HSC/MPP telomere length declined by 30 bp/yr. To interrogate changes in HSC population structure with age, I used the pattern of unique and shared mutations between the sampled cells from each individual to reconstruct their phylogenetic relationships. I found that haematopoiesis in adults aged <65 was polyclonal, with high indices of clonal diversity. In contrast, haematopoiesis in individuals aged >75 showed profoundly decreased clonal diversity. In each elderly subject, 30-60% of haematopoiesis was accounted for by 12-18 independent clones, each contributing 1-34% of blood production. Most clones had begun their expansion before age 40, but only 22% had known driver mutations.

I used the ratio of non-synonymous to synonymous mutations (dN/dS) to identify any excess of non-synonymous (driver) mutations in the dataset. This genome-wide selection analysis estimated that the set of 300 - 400 HSC/MPPs sampled from each adult individual harboured around 100 driver mutations, over 10-fold higher than the number of known drivers we could identify. Novel drivers affected a wider pool of genes than identified in blood cancers.

Simulations from a simple model of haematopoiesis, with constant HSC population size and constant acquisition of driver mutations conferring moderate fitness benefits, entirely explained the abrupt change in clonal structure observed over the age of 70. By old age the majority of HSCs harbour at least one driver mutation. Our data supports the view that dramatically decreased clonal diversity is a universal feature of haematopoiesis in elderly humans, underpinned by pervasive positive selection acting on many more genes than currently known.

Finally, I also sequenced haematopoietic progenitor cells from individuals exposed to a wide range of chemotherapeutic agents. I was able to identify an increased mutation burden associated with a number of chemotherapeutic agents, including platinum and alkylating agents, some of which conferred thousands of excess mutations. There was a wide variation in mutation burden conferred from agents within the same class, meaning that there are potential patient benefits from switching drugs in commonly used regimens. I show that chemotherapy given in childhood can profoundly impact clonal dynamics in later life.

Table of Contents

Preface

The analysis presented in this thesis would not have been possible without help and advice from a large number of individuals. Here, I outline the significant contributions that others have made to the work. I also highlight contributions at relevant points through the text. The experimental design for the work presented in **Chapters 3 and 4** was formulated by myself, Peter Campbell and Elisa Laurenti. The experimental design for the work presented in **Chapter 5** was formulated by Mike Stratton and Peter Campbell, and carried out as part of the CRUK Grand Challenge Mutographs project.

Chapter 1: Introduction

Sections of the text and **Fig. 1.2** are drawn from a review article I wrote with Jyoti Nangalia and Anthony Green, titled 'Clonal approaches to understanding the impact of mutations on haematologic disease development'[1]. The paragraphs on 'mechanisms of ageing' and 'clonal haematopoiesis' are drawn from the manuscript containing work from **Chapters 3 and 4**[2].

Chapter 2: Materials and methods

The majority of this chapter is drawn from the Supplementary Methods section of the manuscript containing work from **Chapters 3 and 4**[2]**.** I wrote this myself, other than the specific sections highlighted below.

- 'Structural variant calling' - written by Hyunchul Jung.
- 'Filtering at the colony level' – contribution by Michael Spencer Chapman
- 'Validation of the phylogenies' – contribution from Michael Spencer Chapman who also performed the benchmarking analysis presented in **Fig. 2.12**.
- 'Using rsimpop to simulate HSC populations' – written by Nick Williams.
- 'HSC population size modelling' – contribution from Kevin Dawson.
- 'dN/dS' analysis' – contribution from Peter Campbell.
- 'Driver mutation acquisition rate estimation' –contribution from Peter Campbell.
- 'Y loss analysis' – contribution from Peter Campbell.
- 'Modelling positive selection in HSC population'- contribution from Kevin Dawson.

- 'Phylofit estimation of selection coefficients' – written by Nick Williams who also performed the phylofit benchmarking analysis presented in **Fig. 2.13**.

Chapter 3: Age-related change in haematopoietic stem cells

I performed all the laboratory work for the work presented in this chapter, with advice from Nicole Mende, Emily Calderbank and Elisa Laurenti. I also undertook all the data curation, figure generation and data analysis, other than the structural variant calling, which was carried out by Hyunchul Jung and LOY variant calling which was carried out by Tom Mitchell. I benefitted greatly from scripts and functions for custom variant filtering and tree building and plotting developed by Mike Spencer Chapman[3] and Nick Williams[4], which I adapted for this work. The HSC population size modelling I performed was completely reliant on the R package *rsimpop*, an HSC population simulator, which was developed by Nick Williams, and a method of posterior predictive checking developed by Kevin Dawson.

Chapter 4: Mechanisms underlying change in HSC population structure with age

As per **Chapter 3**, I performed all the laboratory work to generate the data used in the analysis presented here. I also undertook all the data curation, figure generation and data analysis, other than as outlined below. Again, the modelling I performed was reliant on the R package r*simpop* developed by Nick Williams and the novel method of posterior predictive checking developed by Kevin Dawson. Nick Williams also developed an algorithm, *phylofit*, which I used to estimate driver fitness effects from the pattern of coalescent events in a clade. Fede Abascal and Inigo Martincorena provided useful advice on running *dNdScv* and helped with the validation presented in **Chapter 2**. David Spencer performed the analysis of AML genomes to identify variants in *ZNF318* and *HIST2H3D*.

Chapter 5: Impact of chemotherapy on the haematopoietic system

All laboratory work for this chapter was performed by a research assistant, Anna Clay, under my supervision. I performed all the data curation and data analysis. Scripts to run HDP and assign signatures to trees were adapted from those written by Tim Coorens.

Chapter 6: Discussion

Conclusions drawn are either entirely my own or have been developed through discussion with other lab members and my supervisors Peter Campbell, Elisa Laurenti and Mike Stratton. Some sections are drawn in part from the 'Reviewer's comments' to the manuscript based on **Chapters 3 and 4**[2].

Acknowledgements

First and foremost, I would like to thank my outstanding PhD supervisors Peter Campbell and Elisa Laurenti. They gave me a fantastic project to work on and provided me with the advice and support I needed for it to come to fruition. I have learnt an enormous amount from their different approaches and skill sets – both are incredibly inspirational and have made the perfect supervisory team. I am also hugely indebted to Mike Stratton, who supervised the work in **Chapter 5**, as well as Jyoti Nangalia and Anthony Green who took a supervisory interest in this work and contributed to many helpful discussions. From a career perspective, the last 4 years have definitely been the most exciting and interesting of my life and I will be forever grateful for the opportunity to undertake this work.

Secondly, there are many other individuals who have provided help, advice and support along the way. Some have contributed very directly, as outlined in the preface. In particular, Michael Spencer Chapman, Nick Williams and Kevin Dawson have made many invaluable contributions, without which the analysis presented would not have been possible in its current form. Nicole Mende and Emily Calderback were my mentors on the laboratory side, and I am hugely grateful for their help. Hyunchul Jung kindly undertook the structural variant calling, Tom Mitchell the LOY variant calling and David Spencer the analysis of AML genomes. Many others provided help and discussion along the way including Tim Coorens, Grace Collord, Heather Machado, Henry Lee-Six, Megan Davies, Daniel Hayler, Aditi Vedi, Carys Johnson, Margarete Fabre, Federico Abascal, Lori Kregar, Joe Lee, Phil Robinson, Alex Cagan, Sarah Moody, Yichen Wang, George Vassiliou, Joanna Baxter, Laura Humphreys, Sam Behjati, Inigo Martincorena, Michael Stratton and David Kent. The work presented in **Chapter 5** is in a large part thanks to Anna Clay, a superb research assistant, who performed all the laboratory work and is now undertaking a PhD of her own.

Thirdly, I would like to thank all those who kindly gave or helped me acquire samples, including in particular Krishnaa Mahbubani, Kourosh Saeb Parsy, Krishna Chatterjee and Joanna Baxter. I am grateful to the Cambridge Biorepository for Translational Medicine (CBTM) and Cambridge Blood and Stem Cell Biobank (CBSB). The Cancer, Ageing and Somatic Mutation (CASM) laboratory team, in particular Laura O'Neill and Kirsty Roberts, have been

incredibly helpful in ensuring safe transit of the samples through the Sanger pipelines. In addition, the NIHR Cambridge BRC Cell Phenotyping hub supported all the cell sorting and flow cytometry work.

Fourthly, I would like to thank my family for their encouragement over the last four years, for bearing with me working at odd times, and for taking an interest in the results. I am particularly grateful to my husband Tom, and my parents for all their help and support over the many years that have led to this point.

Finally, I acknowledge funding from Wellcome and the Harrison Foundation, without which work on such a large scale could not have been performed.

List of figures

Main figures

Appendix figures

List of tables

Main tables

Supplementary tables

List of abbreviations

ABC	Approximate Bayesian computation
AML	Acute myeloid leukaemia
BM	Bone marrow
BP	Base pairs
CB	Cord blood
CC	Colon cancer
CNA	Copy number aberration
DDR	DNA damage response
DLBCL	Diffuse large B cell lymphoma
DNA	Deoxyribonucleic acid
FL	Follicular lymphoma
HD	Hodgkin's disease
HDP	Hierarchical Dirichlet Process
HPC	Haematopoietic progenitor
HSC	Haematopoietic stem cell
HSPC	Haematopoietic stem and progenitor cells
LC	Lung cancer
LOY	Loss of Y
LT	Long term
MAD	Mean absolute deviation
MDS	Myelodysplastic syndromes
MNC	Mononuclear cell
MPP	Multipotent progenitor
NA	Not applicable
NB	Neuroblastoma
NHL	Non-Hodgkin's lymphoma
PBS	Phosphate buffered saline
PPC	Posterior predictive check
SBS	Single base substitution
SD	Standard deviation

SNP Single nucleotide polymorphism

SNV Single nucleotide variant

SPL Spleen

ST Short term

SV Structural variant

VAF Variant allele fraction

WGS Whole genome sequencing

Chapter 1: Introduction

This thesis aims to answer important questions about how the haematopoietic system changes with age, both at the level of single haematopoietic stem cells (HSCs) and at the level of the entire HSC population. In addition, the impact of chemotherapy in perturbing or altering age-related changes in HSC mutation burden and population structure will be addressed.

The introduction to the thesis will cover useful background under the following headings:

1. Haematopoietic stem cells.
2. History of clonal approaches used in the study of haematopoiesis.
3. Mechanisms of ageing.
4. Features of haematopoietic stem cell ageing.
5. Haematopoietic stem cell population dynamics.
6. Impact of chemotherapy on haematopoietic stem cells.

The chapter will conclude by outlining the aims of the thesis.

1.1 Haematopoietic stem cells

The haematopoietic hierarchy

Haematopoiesis is a complex process that supports the immune, oxygen carrying and haemostatic functions of blood through the coordinated production of hundreds of billions of mature blood cells each day. HSCs reside at the apex of the haematopoietic lineage tree. The traditional haematopoietic hierarchy depicted HSCs as a homogenous population that underwent lineage commitment through a series of stepwise changes, and discrete intermediate cell states. However mounting evidence now supports a 'continuous' model, in which the phenotypic HSC compartment is molecularly and functionally diverse and there is a continuum of transitionary cell states leading to all mature cell blood cell types, as illustrated in **Figure 1.1**. Support for this 'continuous' model of haematopoiesis come from scRNA

datasets[5] and *in vivo* barcoding approaches in both mice[6,7] and humans[8] (in the setting of gene therapy).

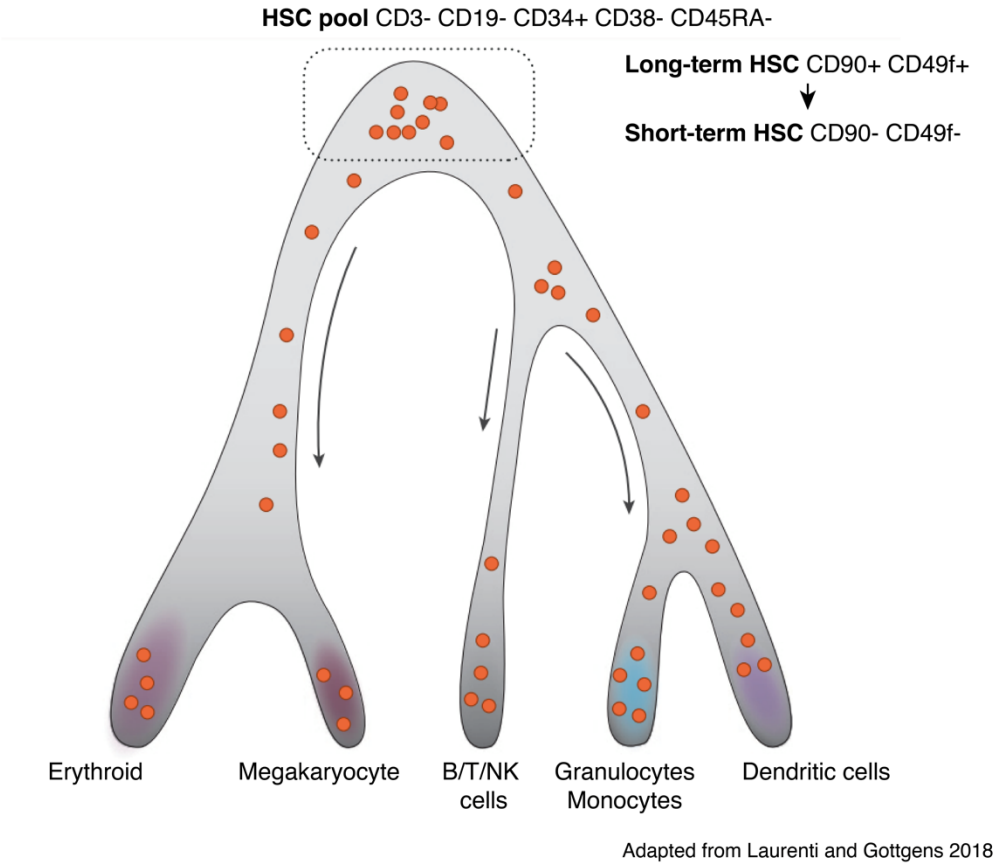


Fig. 1.1| Continuous model of haematopoiesis. Haematopoietic stem cells sit in the HSC pool at the top of the haematopoietic hierarchy. The 'HSC pool' is molecularly and functionally heterogeneous and contains long-term HSCs, short-term HSCs and multipotent progenitors (MPPs). Differentiation towards discrete cell types occurs along a continuum of transitionary states.

Haematopoietic stem cell definitions

The classic definition of a haematopoietic stem cell is a cell that is able to give rise to all mature blood cell lineages and self-renew. The gold standard approach to defining human HSCs uses *in vivo* functional assays. Long-term HSCs are defined as having sustained repopulation capacity in primary and secondary xeno-transplantation. In contrast short-term HSCs and multipotent progenitor populations (MPPs) are defined as having only transient multilineage engraftment in primary animals and no capacity for secondary transplant[9]. However, it is being increasingly recognized that this xenotransplant approach to defining

human HSC potential has potential issues, not least the impact of transplantation into another species with the altered niche microenvironment and signaling that this entails. In addition, *in vitro* culture assays are able to demonstrate multipotential output in a subset of phenotypic HSCs. These culture assays are however only able to 'rule in' stem cell potential, and cannot exclude the possibility of additional stem cell potentials *in vivo* or in the context of alternative culture conditions. Both the xenotransplant and *in vitro* culture approaches are highly subject to technical limitations.

Recent work has helped refine the immunophenotypic definition of the HSC compartment through the use of surface markers to prospectively isolate subsets within this pool. The marker CD34 alone identifies >99% of human haematopoietic stem and progenitor cells (HSPCs) but true long-term (LT) HSCs make up <1% of the CD34+ population. Work in the 1990s identified additional antigens (CD90+, CD38-, CD45RA-), that enriched for self-renewal capacity[10–12]. More recently, a CD49f ^{hi}/CD90 ^{lo} gating strategy has been shown to isolate LT-HSCs in cord blood at a purity of 1 in 10.5[13]. Despite these improvements in isolation strategies, work on human HSCs is still confounded by the heterogeneity of cells present even in the tightest immunophenotypically defined populations. In addition, most studies in the field have been limited to analysis of cord blood (CB) HSCs, which are unlikely to be representative of the bone marrow (BM) or peripheral blood (PB) HSC pool populations in adults.

Unique features of the haematopoietic system

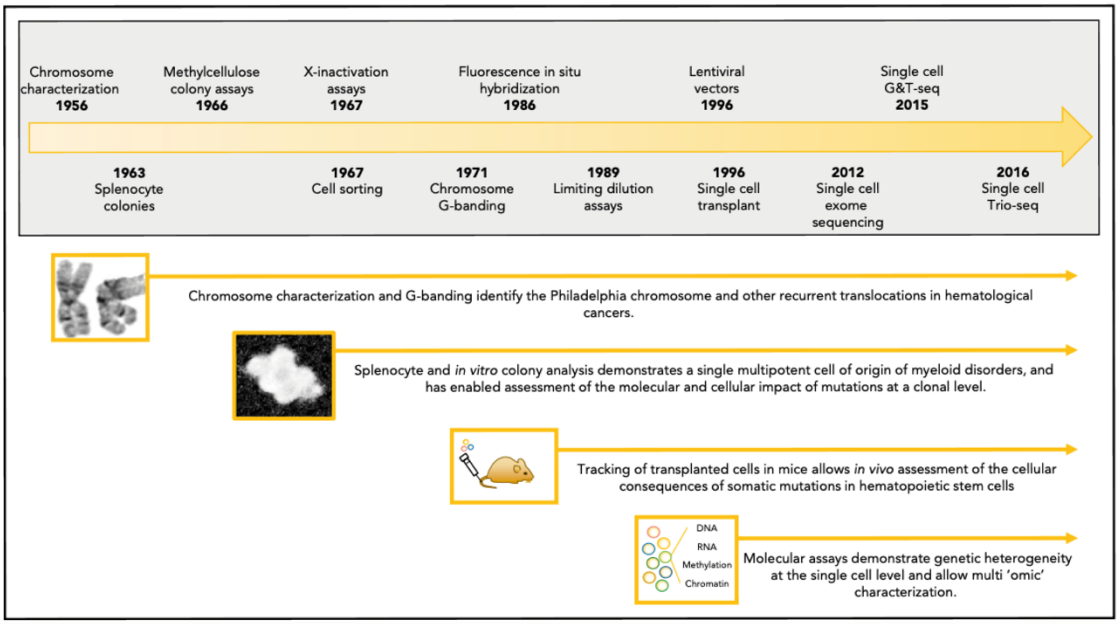
As compared with stem cells in other tissues, HSCs are unique in having minimal spatial constraints. In adults, HSCs are regulated by a distinct hypoxic environment in the BM, while a minor fraction circulates through the PB and other organs. Current evidence supports the view that HSCs within the BM and PB are well mixed, with no significant spatial segregation of malignant or normal clones evident within the bone marrow, at least over timescales greater than a few years. Variant allele fractions (VAFs) of pathogenic mutations in patients with myelodysplastic syndromes (MDS) are highly concordant between BM and PB samples[14], and the same has been observed when looking at VAFs of non-pathogenic somatic mutations in one normal individual[15].

Another unique feature, is the ease with which HSPCs can be clonally expanded *in vitro*. This is one of a number of 'clonal' approaches, discussed in more detail below, that has been used in the study of haematopoiesis.

1.2 History of clonal approaches used in the study of haematopoiesis

Early clonal approaches to understanding haematological malignancy

Interrogation of haematopoiesis at the clonal level was pioneered in blood and has a rich history spanning over 50 years (**Fig. 1.2**). The identification of the Philadelphia chromosome as a clonal abnormality in cells from patients with chronic myeloid leukaemia was seminal in implicating the first specific genetic mutation as a cause of cancer[16]. Subsequent cytogenetic techniques, such as Giemsa-banding of metaphase chromosomes and fluorescence in-situ hybridisation, identified a range of chromosomal lesions present in many leukaemias and lymphomas[17].



From Nangalia, Mitchell and Green 2019

Fig. 1.2| Timeline of haematopoietic clonal experimental approaches. Clonal approaches illustrated include chromosome characterization, hematopoietic colony assays, transplantation studies, and sequencing-based techniques. Major milestones in the development of these approaches are shown in the timeline. G&T-seq, genome and transcriptome sequencing; Trio-seq, single-cell triple omics sequencing.

Studies of X-chromosome inactivation patterns have been another cornerstone of our understanding of the clonal origin of hematopoietic neoplasms. Expression studies of X-linked glucose-6-phosphate dehydrogenase (*G6PD*) in females genetically heterozygous for the *G6PD* locus were first used to study cancer in patients with leiomyomas. This work identified

that tumour cells expressed only one *G6PD* allele, suggesting their unicellular, or clonal, origin[18]. Similar findings were made in females with lymphoma[19] and chronic myeloid leukaemia[20], and in the latter, the presence of monoallelic expression of *G6PD* across the different differentiated hematopoietic cell types further suggested that the tumour arose from a multi-potent stem cell[20]. Such studies in polycythaemia vera, a disease not previously recognised as being neoplastic, also established it as a clonal disorder arising in a multipotent progenitor or stem cell[21]. In addition, the recognised skewing of X-activation patterns often found in blood from elderly females paved the way for the subsequent discovery of age related clonal haematopoiesis[22].

Clonal approaches to understanding haematopoiesis

Clonogenic assays involving the *in vitro* expansion of single myeloid progenitors were developed around the same time as studies of X inactivation[23], and helped define the hematopoietic cellular hierarchy. Diluted bone marrow or peripheral blood derived mononuclear cells plated in semi-solid media and cultured in the presence of colony stimulating growth factors, result in the growth of distinct individual 'colonies'. Each colony comprises a cluster of differentiated cells derived from a single progenitor cell. Using this approach, many myeloid diseases were dissected at the colony level in the 1970s. More recently, colonies of cells can also be grown in liquid culture within individual wells seeded with single cells of interest. The ability to isolate specific haematopoietic cell populations using flow cytometry, and the availability of an array of *in vitro* culture conditions, has since allowed the growth of colonies from specific starting hematopoietic populations and finer resolution of genotype-phenotype relationships[24].

Growth of splenocyte colonies in the 1960s using donor cells traceable in recipient mice identified for the first time that different myeloid cell types are derived from a common multipotent hematopoietic progenitor cell[25]. Technical advances enabling single-cell transplantation assays are able to assess stem cell fitness and engraftment potential of individual cells harbouring specific mutations, albeit in the environmental context of an irradiated recipient[26]. Clones derived from human HSCs can also be characterised following xenotransplantation into mice[27] using endogenous markers (for example specific genetic mutations or rearrangements) or by introducing markers[28] (for example lentiviral vectors[29],

genetic barcodes[30]). More recently, CRISPR scratchpads have been used for single-cell clonal tracing in zebrafish to dissect the embryonic relationships between adult cell types[31].

Newer technologies applied to the haematopoietic system include single-cell whole-exome sequencing[32,33], and G&T-seq[34], which assesses both genomic and transcriptomic information in single cells. However these approaches suffer from both false-negative and false-positive mutations due to the paucity of genomic material and requirement for DNA amplification[32,33]. Due to these and other technical limitations these methods are yet to be applied widely.

Colony-based approaches circumvent the issues with single-cell technologies, by allowing the generation of large amounts of clonal starting material from single HSPCs. This allows highly accurate exome and whole-genome sequencing of somatic mutations present in the seeding cell. The work presented in this thesis utilises whole genome sequencing (WGS) of single-HSC derived colonies to allow an unbiased and accurate interrogation of molecular features of ageing at the level of single HSCs, and changes in HSC clonal dynamics with age.

1.3 Mechanisms of ageing

Theories of ageing

The age-related mortality curve for modern humans is an outlier across the tree of life, with an abrupt increase in mortality after the average lifespan[35], leading to surprisingly low variance in age at death[36]. Studies of ageing at the cellular level have demonstrated that accumulation of molecular damage across the lifespan is gradual and lifelong, including somatic mutation[37–39], telomere attrition[40–42], epigenetic change[43] and replicative stress (or stem cell exhaustion)[44,45]. The cell intrinsic accumulation of molecular damage has been proposed to underlie ageing phenotypes[46]; however it remains unresolved how such gradual accumulation of molecular damage can translate into an abrupt increase in mortality after the age of 70 years.

Some evidence also supports the view that cell extrinsic factors and microenvironmental changes are implicated in ageing phenotypes. In particular the concept of ‘inflamm-ageing’, whereby increased inflammatory stimuli with age contribute to age-related loss of function

in organ systems has been widely postulated[47]. There is also limited evidence that an aged bone marrow niche is able to cause HSC dysfunction, and that extrinsic signals can rejuvenate 'aged' HSCs[48]; although the paper reporting these findings has now been retracted casting doubt on its validity.

The work presented in this thesis focuses on patterns of somatic mutation accumulation and telomere attrition in single HSCs with age, as these are the two types of molecular damage that can be accurately assessed using the colony-based WGS approach.

Somatic mutation accumulation

Every cell division throughout life, starting from the first division of the fertilised egg, requires the accurate replication of the entire genome, which in humans comprises 3 billion nucleotide base pairs. Mammalian DNA replication and repair systems, whilst highly complex and precise, are not infallible. In addition, ongoing exposure to endogenous and extrinsic DNA damaging insults inevitably results in the acquisition of somatic mutations in individual cells[49].

Thus, the DNA composition of cells within a tissue can be compared to a fine mosaic, each distinct tile of which is akin to an individual cell that differs from its neighbour by virtue of its unique catalogue of somatic mutations. The vast majority of such mutations occur in non-coding regions of the DNA and are believed to have a neutral effect on cellular fitness[50]. However, mutations can lead to positive selection and clonal expansion if they land in genomic regions, for example those with oncogenic potential, that result in cellular phenotypes which enhance fitness with respect to competing normal cells (termed 'driver' mutations). Clonal expansion provides a reservoir for the acquisition of further driver mutations, and ultimately carcinogenesis. There is also the theoretical possibility that some mutations can lead to negative selection, although estimates of this in cancer evolution show the effect to be negligible, with purifying selection almost absent outside homozygous loss of essential genes[50].

Somatic mutations in cells are acquired in several ways. Cell-extrinsic mutagens include chemicals (such as tobacco and aflatoxin), ionizing radiation and ultraviolet light. Additionally, DNA is damaged via cell-endogenous exposures to reactive oxygen species, inadequate

function of DNA repair enzymes, abnormal activity of DNA-editing enzymes, and activity of viruses and retrotransposons[1]. Interestingly, haematopoietic cancers and some paediatric brain tumours carry the smallest number of somatic mutations across all human cancers[51], suggesting that, compared to many other tissues, HSCs are relatively well protected from this mutagenic onslaught.

One key study performed whole-exome sequencing on hematopoietic colonies grown from cord blood or from individuals of varying ages, and identified a linear increase in mutation burden with age[52]. The number of somatic mutations found in acute myeloid leukaemia (AML) in their study was close to that expected in a normal individual of the same age. These data provided two important insights: first, that most mutations detected in bulk AML samples represent those that have accumulated with age and were present prior to malignant transformation; and secondly, that background somatic mutation acquisition in HSCs can be viewed as a molecular clock, with the total number of mutations present reflecting the age of the individual.

These findings have since been corroborated by whole genome sequencing studies in normal tissues. In blood, a small number of single-cell derived blood progenitor colonies underwent WGS from individuals under the age of 65, with the findings in keeping with a linear acquisition of mutations to this age[53]. Mutation burdens based on WGS have also found lower mutation burdens in blood stem and progenitor cells than in other normal tissues sequenced to date, including lung[54], colon[55], bladder[56] and endometrium[57], in keeping with the observation of low mutation burdens in myeloid blood cancers.

Telomere attrition

Due to the fact that DNA polymerases are unable to completely replicate the terminal ends of DNA molecules, telomeric regions of chromosomes are eroded with each cell division. Replication of the terminal ends of telomeres requires a specialised DNA polymerase, known as a telomerase. However, the majority of human somatic cells (including HSCs) do not express telomerase, leading to the progressive attrition of telomeric regions over life[42]. *In vitro* work on human HSCs has shown they lose 30-100bp of telomeric DNA per cell division[58].

Loss of telomeric DNA eventually leads to cellular senescence, a state of arrested cell division, and / or apoptosis due to the activation of DNA damage response (DDR) pathways. DDR pathways are activated through exposure of the blunt ends of chromosomal DNA, which are recognised as double strand breaks. Telomere attrition is thought to explain the limited proliferative capacity of cultured cells *in vitro*, the so-called Hayflick limit of between 40-60 cell divisions before entering a senescence phase[59]. Protection of cells from senescence and apoptosis triggered by DNA damage response pathways also requires that telomeric DNA is bound to nucleoprotein 'shelterin' complexes[60].

Another rare outcome of telomere crisis it that of massive genomic rearrangement induced by naked chromosome ends triggering DNA repair. Breakage-fusion-bridge cycles are common in this situation and chromothripsis, the shattering and sticking together of large chromosomal regions, is also likely to be triggered by critical telomere attrition. These rare phenomena are of biological importance as they can be associated with cancer progression[61].

The potential importance of telomere attrition for ageing phenotypes is highlighted by the observation that both telomerase deficiency and shelterin mutations can underlie dyskeratosis congenita and aplastic anaemia, as well as accelerated ageing phenotypes in other tissues[40,41,62].

1.4 Features of haematopoietic system ageing

The human haematopoietic system displays a number of clinically detectable phenotypes with ageing. These include reduced regenerative capacity and cytopenias, particularly in the face of insults such as infection, blood loss or chemotherapy; reduced immune function, and an increased risk of both clonal haematopoiesis and myeloid malignancy.

Reduced regenerative capacity

Many elderly individuals have mild cytopenias of unknown aetiology, representing a reduction in the ability of the haematopoietic system to produce mature blood cells. A new WHO diagnostic category, idiopathic cytopenias of uncertain significance (ICUS)[63,64], has been developed for those individuals with any of haemoglobin < 110 g/L, platelets < $100x10^9$/L or

neutrophils < $1.8x10^9$/L. Additional diagnostic criteria for ICUS include the presence of cytopenias in the absence of dysplasia and driver mutations in known myeloid malignancy genes. The cellular mechanisms underlying the cytopenias in these individuals remains unclear but a proportion go on to develop haematological malignancies, most commonly MDS.

In mice it has been shown that while the number of phenotypic HSCs increases with age, aged HSCs have significantly reduced self-renewal capacity in serial transplantation compared to young HSCs [65–67]. There is also increased functional heterogeneity in the aged HSC pool, with occasional aged HSCs having a potent or 'young' phenotype[65,68,69].

Loss of immune function

Loss of immune function is a well-described feature of ageing[70]. In humans, loss of immune function with age results in both an increased risk of viral and bacterial infections and in the increased incidence of autoimmune disease. One contributing factor to the decline in the adaptive immune system, is an increase in myeloid-biased HSCs with age, which functionally show decreased lymphoid cell output. This phenomenon has been described in both mice[71] and humans[72]. A single-cell transplantation study has identified massively increased numbers of myeloid-restricted progenitors in the phenotypic HSC compartment of aged mice[73]. In addition, a label retention study using pulsed histone 2B-green fluorescent protein[74] showed that functionally multi-potent HSCs had undergone very few cell divisions (label retaining), while HSCs that had undergone many cell divisions (loss of label) show myeloid-restricted potential. There is also experimental support for increased HSC cell division caused by the stimulus of infection and inflammation contributing to the skewing of lineage outputs seen in mice[75]. Together these studies suggest that aged HSCs have altered differentiation potential that likely plays a role in the decline of the adaptive immune system.

Clonal haematopoiesis

One area of recent research interest has been 'clonal haematopoiesis', an age-related feature of the haematopoietic system identified in the last decade. Sequencing of blood samples from population cohorts has revealed an age-related increase in acquired mutations in genes that

cause myeloid neoplasms[22,76–79], known as 'driver mutations'. Clonal haematopoiesis reaches 10-20% prevalence[22,76–79] or even higher (using error corrected sequencing methods)[80] after 70 years. However, the sum of driver mutations in normal individuals typically accounts for only a small fraction of total haematopoiesis (<5% cells). The most commonly mutated genes driving clonal haematopoiesis in normal elderly individuals include the epigenetic modifying genes *DNMT3A*, *TET2* and *ASXL1*, and the splicing genes *SF3B1* and *SRSF2*. Mutations in *DNMT3A* and *TET2* increase in frequency roughly linearly with age, while splicing gene mutations are almost never seen in individuals aged < 60 but are found in a relatively large proportion of the very elderly[81]. The mechanisms underlying this difference in prevalence of different types of gene mutation with age remain unclear.

Some elderly individuals have evidence of clonal expansions even in the absence of known driver mutations[82–85]. The most comprehensive of these studies by Zink *et al*[82] analysed a population cohort of blood WGS data, identifying clonal expansions, based on the number of callable somatic mutations, in just over 1% of individuals. In these cases of identifiable clonal expansions, mutations in 18 known myeloid malignancy genes could only account for 18% of the putative clonal expansions. However, deep sequencing of bulk blood samples, as used in the studies that have identified clonal expansions in the absence of known driver mutations to date, struggles to elucidate clonal relationships among cells and is insensitive to mutations present in <1-10% of blood cells.

Due to the limitations of previous studies of clonal haematopoiesis described above, an unbiased, high-resolution model for the clonal dynamics of human haematopoiesis with ageing was lacking at the outset of this project. Whole-genome sequencing of colonies grown from single cells circumvents the limitations of bulk sequencing and is the approach taken in this thesis[53,86,87].

1.5 Haematopoietic stem cell population dynamics

Despite some advances in our understanding of human HSC ageing, parameters relating to how HSC population dynamics change with age remain relatively poorly defined. These parameters include changes in absolute HSC number and changes in HSC generation time (the time in years between symmetric self-renewal divisions).

Estimates of HSC population size and generation time

A number of approaches have been used previously to assess the number of HSCs throughout life. Analyses based on progressive skewing of X-chromosome inactivation patterns in women estimated the number of human stem cells in the adult steady state at around 11,000 with an average replication rate of once every 40 weeks[88]. However, the finding that skewing of X-chromosome inactivation is frequently driven by the emergence of a single clone carrying a driver mutation[22] raises doubt about the underlying statistical model in this type of approach. Other studies have used the patterns of telomere shortening to infer stem cell dynamics[89] showing that the HSC population appears to expand during adolescence and be maintained during adulthood, but there is limited resolution to make inferences beyond this.

More recently, Lee-Six *et al* have used somatically acquired mutations in the HSCs and haematopoietic progenitors from a healthy 59 year old man to infer the phylogenetic relationship between HSC lineages[86]. HSPC colonies were grown *in vitro* to generate enough clonal material for whole genome sequencing. The branching pattern of the phylogenetic tree generated was used to estimate current and historical HSC population size and targeted recapture of mutations in peripheral blood to estimate the contribution of each branch to mature cell production (**Fig. 1.3**). The total number of adult HSCs was estimated to be much larger than previously thought (between 44,000 and 215,000), with each HSC undergoing a symmetrical division every 2-20 months (generation time 0.16 – 1.6 years). Another recent study that took a completely orthogonal approach, using VAFs of *DNMT3A* R882 in large cohorts, also estimated HSC population size to be in the same range as the Lee-Six *et al* estimate. It found the most credible value for human HSC $N\tau$ (product of population size *N*

and generation time τ) to be approximately 100,000. Other estimates of human HSC generation time in the literature range between 0.6 and 6 years[42,86,88].

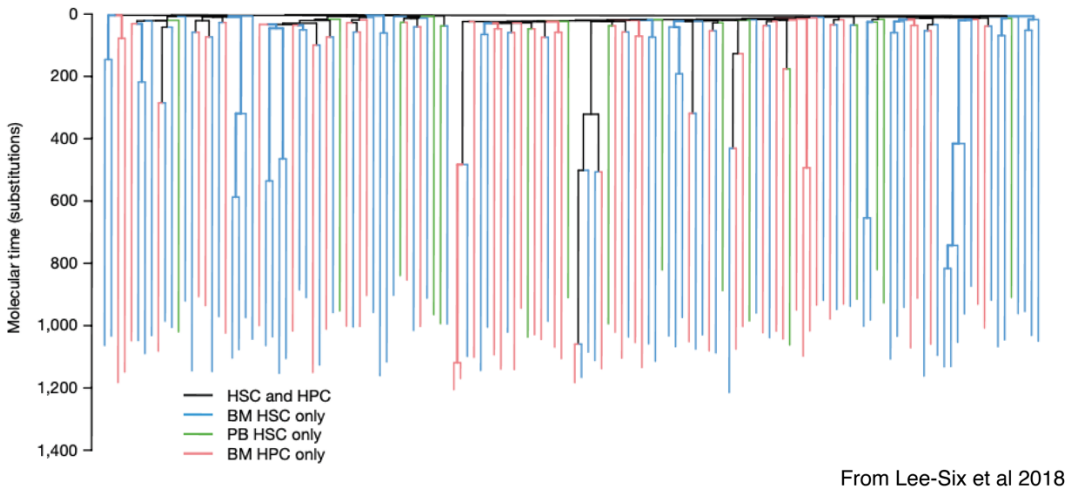


Fig. 1.3| First HSPC phylogeny created from the bone marrow of a normal 59 year man. Phylogeny of 140 single haematopoietic stem and progenitor cells showing the relationship between cell types. At each tip of the tree is a colony. Branches connect colonies to each other to form a family tree. Branch lengths are proportional to the number of somatic mutations. Branches are coloured according to the phenotype of their descendants. Branches ancestral to haematopoietic progenitor cells (HPCs) are coloured red; branches ancestral to bone marrow-derived haematopoietic stem cells (BM HSCs) are blue; branches ancestral to peripheral blood-derived haematopoietic stem cells (PB HSCs) are green; branches ancestral to both stem and progenitor cells are coloured black.

Lee-Six *et al* were able to make estimates of HSC population size and generation time using a phylodynamic approach that was first pioneered by pathogen epidemiologists[90]. The basis of this approach is discussed below.

The phylodynamic approach

Phylodynamics is the study of how different processes, such as fluctuations in populations size and positive selection, act and interact to shape phylogenies, or family trees. One fundamental tenet of phylodynamics is that the frequency of branching events (known as 'coalescences') in a phylogeny created from a random sample of a population is defined by $N\tau$, where N is the population size and τ is the generation time. This means that the same phylogeny could be obtained from a population of 100,000 with a generation time of 1 year ($N\tau$ = 100,000) and a population of 25,000 with a generation of 4 years (again $N\tau$ = 100,000).

When applied to somatic cell populations, coalescent or branching events represent historic symmetric self-renewal divisions.

It has been shown that in a neutrally evolving population the pattern of coalescent events in a phylogeny created from a random sample of individuals can be used to infer historic population size changes[90]. Specifically, in populations of a constant size (*N*) and generation time *(τ)* there will be more coalescent events per generation (which define individuals that are related) observed when *N* is small compared to when *N* is large. The reason for this difference in phylogenies from small and large populations can be understood by imagining the predicted phylogeny obtained from sampling 10 random individuals from a population of 50 individuals, compared to sampling the same number from a population of 500. We would expect to have a higher chance of sampling siblings and cousins from the smaller population than the larger, which manifests as more coalescent events in the phylogeny per generation, or in the case of somatic cells per unit of molecular time (**Fig. 1.4**). This concept can be taken a step further, such that in a population with a fluctuating population size, more coalescent events will be observed in time 'windows' where the population size is small compared to when it is larger.

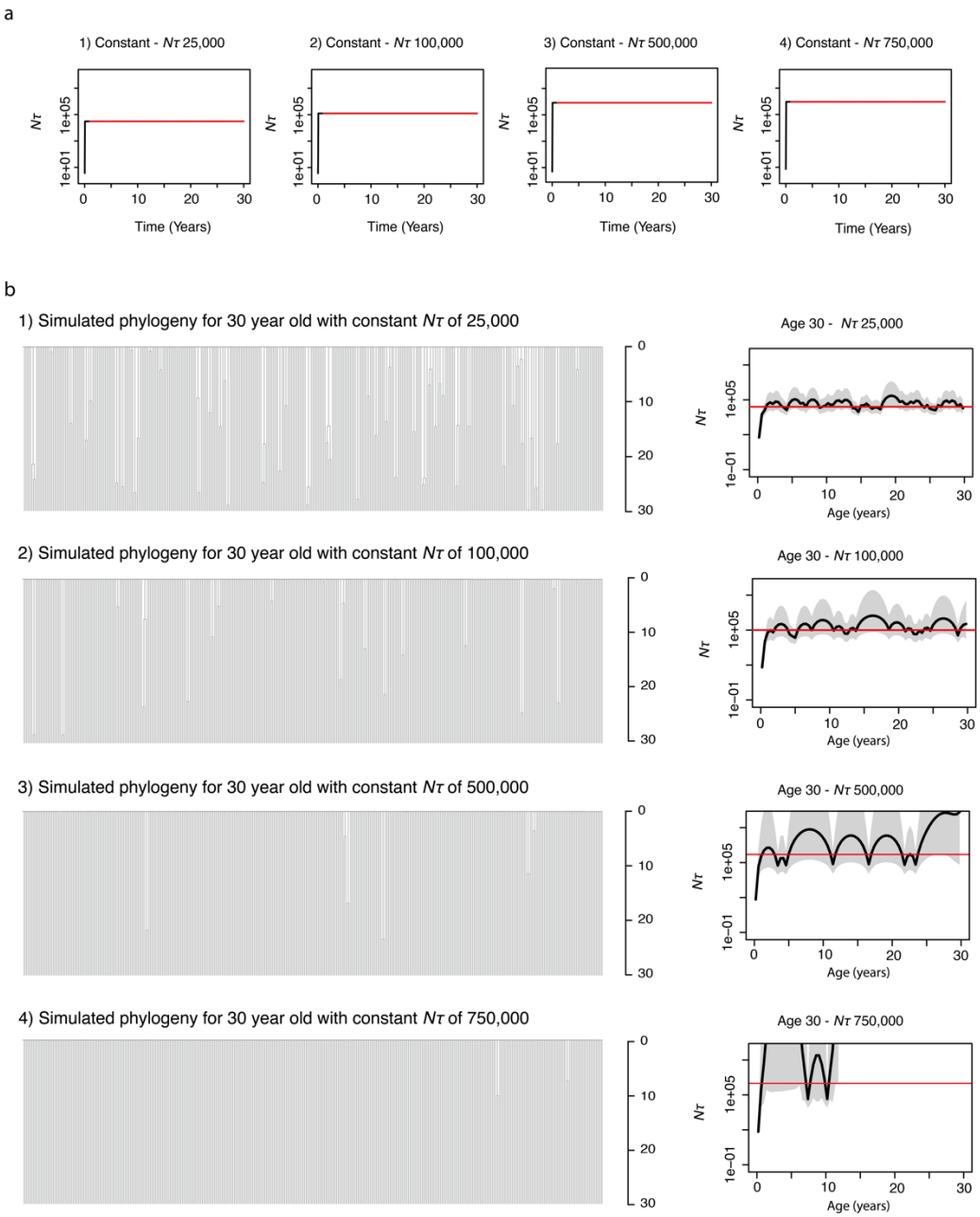


Fig. 1.4| Effect of population size on phylogeny structure. a, Trajectories of $N\tau$ used as input to *rsimpop* for the simulations to create phylogenies in b. Note the Y axis depicting $N\tau$ is on a log scale. **b,** Phylogenies created by randomly sampling 380 cells from the final full simulated population of between 25,000 cells (Phylogeny 1) and 750,000 cells (Phylogeny 4). Phylogenies 1 to 4 are all derived from simulations of the HSC population up to the age of 30 years. Each simulation has an $N\tau$ of 100,000. In all cases $N\tau$ is the same as the population size (N), as the generation time (τ) is 1 year. The *phylodyn* trajectories to the right of each simulated phylogeny use the pattern of coalescent events to recover the input trajectories for $N\tau$.

The action of genetic drift in a population means that a proportion of lineages are lost stochastically per unit time. Therefore, when somatic stem cell populations during homeostasis are considered, the older the individual the more coalescent events will be observed per unit time. Genetic drift is accounted for in phylodynamic models such as *phylodyn*[90].

The action of positive selection in a population will alter the pattern of coalescent events in a phylogeny if it results in detectable clonal expansion. This means that inferences of population size and historic population dynamics are only valid in populations that do not have evidence of high levels of positive selection. The approach also relies on the random sampling of cells within the population.

Application of phylodynamics to stem cell populations

When applied to stem cell populations, *N* is the number of stem cells in the population and τ is the generation time. Stem cells can divide in three distinct ways. The first is a symmetric self-renewal division that creates 2 daughter stem cells, so increasing the stem cell population and being the equivalent of stem cell birth. The second is a symmetric differentiation division that results in 2 differentiated daughter cells, which is the equivalent of stem cell death. The final type is an asymmetric division, which produces one stem cell and one differentiated cell and therefore does not alter the size of the stem cell population. It can be seen that only the symmetric self-renewal division results in a daughter progeny that increases the size of the stem cell population. From the phylodynamic perspective stem cell generation time is therefore defined as the time between symmetric self-renewal divisions.

Due to the requirement to sample random cells within a population, it is impossible to robustly apply phylodynamic methodology to somatic stem cells in solid organs. However, the haematopoietic system is the one example of a somatic stem cell population that can be randomly sampled, either through peripheral or cord blood sampling, or by taking a large bone marrow sample from multiple bones. The sampling of large volumes of bone marrow (50-80ml) from deceased organ donors provides the additional advantage that these individuals are highly likely to have had high levels of circulating cytokines at the time of

sampling which is known to mobilise HSCs within the bone marrow[91,92]. This makes the haematopoietic stem cell population sampled in these ways the ideal candidate for the application of phylodynamic methods. Nevertheless, interpretation of the phylogenies created by sampling HSCs can be non-intuitive. **Appendix 1** therefore expands on the simulated phylogenies in **Figure 1.4** to aid interpretation of results presented in **Chapter 3**.

1.6 Impact of chemotherapy on the haematopoietic system

The three main ways that chemotherapy might be expected to impact HSCs are 1) Mutagenic effect causing increased mutation burden; 2) Cytotoxic effect reducing HSC population size or 3) Altering the landscape of positive selection on HSCs.

Mutational signatures

Excess mutations attributable to specific mutagenic exposures have been found to produce characteristic 'mutational signatures'[49,93–95], most readily identifiable in WGS datasets. 'Mutational signatures' are patterns of mutagenesis that occur in specific trinucleotide contexts, with some chemotherapeutic agents identified as causing specific signatures in human cancer. Single base substitution (SBS) signatures represent substitutions (C>A, C>G, C>T, T>A, T>C and T>G) considered in the context of the flanking 3' and 5' bases, giving a total of 96 different contexts. The COSMIC database of mutational signatures derived from the PCAWG (cancer) dataset (https://cancer.sanger.ac.uk/signatures/sbs/) contains 60 mutational signatures of which 7 are proposed to be due to chemotherapeutic agents (**Table 1.1**) and many are yet to have an aetiology ascribed.

Table 1.1: Mutational signatures caused by chemotherapy in cancerous tissues

Mutational signature	Exposure
SBS11	Temozolamide
SBS25	Unknown chemotherapy
SBS31	Platinum agent
SBS32	Azathioprine
SBS35	Platinum agent
SBS86	Unknown chemotherapy
SBS87	Thiopurine agent

Two of these signatures could not be attributed to a specific agent, and the others represent only a tiny fraction of the chemotherapeutic agents currently in use. The relatively limited selection of chemotherapeutic agents represented in the COSMIC SBS signature database is likely at least in part due to the fact that the majority of PCAWG cancer samples were taken at diagnosis prior to exposure to any chemotherapy. In addition to the 7 signatures attributable to chemotherapy in the COSMIC database, analysis of myeloma samples post melphalan autograft has identified a putative melphalan signature[96–98] and analysis of metastatic cancer samples has found a SBS17b-like 5FU/capecitabine signature in bowel cancers[99].

Evidence for an impact of chemotherapy on normal tissues has been much less well characterised. There is evidence that the 5FU/capecitabine signature is only found in or close to cancerous bowel tissue, with more spatially distant regions completely spared[100]. However other agents have been found to affect the whole normal organ, for example an SBS25-like signature was found in every normal colonic crypt sampled in an individual previously treated for Hodgkin's lymphoma[101]. No investigation of the impact of chemotherapy in normal blood at the single cell level had been made at the outset of this project, although a recent case report has identified the platinum signature in bulk blood samples of two children treated with platinum agents for neuroblastoma[102].

Impact of chemotherapy on HSC population size and generation time

The impact of chemotherapy on the size of the human HSC population and its generation time has not previously been assessed, and is likely to be dependent on both the chemotherapeutic agent and its dose. Clinically it is known that high dose chemotherapy given in the setting of allogeneic transplant can eradicate the normal HSC population almost completely. However, the impact of lower dose chemotherapeutic regimens on the human HSC population size and generation time is not clear. In mice, the LT-HSC population does not recover to a normal size post inducible depletion, remaining around 10% of normal levels without compromise to downstream haematopoiesis[103]. Murine studies of 5FU treatment have shown chemotherapy results in increased HSC turnover and reduced generation time[104,105]

Impact of chemotherapy on HSC selective landscape

Chemotherapy has been shown to impact the HSC selective landscape, favouring cells able to continue dividing despite DNA damage. Clonal haematopoiesis driven by mutations in the DNA damage response pathway genes *TP53*, *PPM1D* and *CHEK2* is more prevalent in individuals exposed to previous chemotherapy[106–108]. Clones carrying mutations in these genes have also been shown to expand most rapidly when assessment of clone size is made pre-and post-chemotherapy[106]. At least in adult individuals, the vast majority of clones carrying driver mutations identified post chemotherapy are also detectable in pre-chemotherapy samples. This would suggest that in the majority of chemotherapy exposed individuals the effect on selective landscape may be more important than the mutagenic impact of chemotherapy[106].

Risk of secondary malignancy

In a minority of chemotherapy exposed individuals (up to 5-10% depending on the regimen), chemotherapy contributes to secondary haematological malignancy, most commonly therapy-related MDS and acute myeloid leukaemia (AML)[109]. In these cases, it is likely that the mutagenic impact of chemotherapy plays a more important role. One unanswered question relating to secondary malignancies, is why their risk should be highest in the 5-10

years immediately post chemotherapy, dropping back to close to normal levels beyond this time[109].

1.7 Summary

In summary, the haematopoietic system is unique amongst organ systems in being comprised of a spatially well mixed population of stem cells capable of coordinating the production of hundreds of billions of mature blood cells each day. The mechanisms underlying age-related phenotypes in the haematopoietic system, which typically only become clinically apparent over the age of 65-70, remain unclear. The impact of chemotherapy on normal HSCs has also not previously been well characterised. These areas of research are the focus of this thesis, with the specific aims to be addressed in each results chapter outlined below.

1.8 Aims

Aim 1:
Characterise aspects of age-related change in normal human HSCs at the single cell and population level (**Chapter 3**)

Aim 2:
Determine mechanisms underlying changes in human HSC population structure with age (**Chapter 4**)

Aim 3:
Investigate the effect of chemotherapy on HSC age-related changes and population structure (**Chapter 5**)

Chapter 2: Materials and methods

2.1 Introduction

This chapter outlines the methods used for all the work presented in this thesis. **Chapters 3 and 4** present results based on data obtained from samples that form the 'ageing dataset' (**Table 3.1**). **Chapter 5** presents results on data from samples that form the 'chemotherapy dataset', which comprises both samples from chemotherapy exposed individuals and comparator normal individuals (**Tables 5.1 and 5.3**). The normal samples used for comparisons of mutation burden and mutational signatures in the 'chemotherapy dataset' include a subset of those used in the 'ageing dataset', as well as samples from two additional individuals. The methods used for the two different datasets differ slightly and this is signposted where applicable.

Some sections of the methods are almost entirely written by colleagues who have contributed to these parts of the analysis. Where this is the case I have made it clear at the start of the section.

2.2 Samples

Normal individuals

In order to obtain representative data from across the whole human lifespan samples were obtained from three sources: **1)** Stem Cell Technologies provided frozen mononuclear cells (MNCs) from two cord blood samples that had been collected with informed consent, including for whole genome sequencing (catalog #70007). **2)** Cambridge Blood and Stem Cell Biobank provided fresh peripheral blood samples taken with informed consent from two patients at Addenbrooke's Hospital (NHS Cambridgeshire 4 Research Ethics Committee reference 07/MRE05/44 for samples collected pre-November 2019 and Cambridge East Ethics Committee reference 18/EE/0199 for samples collected from November 2019 onwards). **3)** Cambridge Biorepository for Translational Medicine provided frozen bone marrow +/-

peripheral blood MNCs taken with informed consent from six deceased organ donors. Samples were collected at the time of abdominal organ harvest (Cambridgeshire 4 Research Ethics Committee reference 15/EE/0152).

Chemotherapy exposed individuals

Samples from chemotherapy-exposed individuals were obtained from Cambridge Blood and Stem Cell Biobank who provided fresh peripheral blood samples taken with informed consent from a total of 19 patients at Addenbrooke's Hospital (NHS Cambridgeshire 4 Research Ethics Committee reference 07/MRE05/44 for samples collected pre-November 2019 and Cambridge East Ethics Committee reference 18/EE/0199 for samples collected from November 2019 onwards).

2.3 Isolation of MNCs from fresh peripheral blood samples

Mononuclear cells (MNCs) were isolated using lymphoprep^{TM} density gradient centrifugation (STEMCELL Technologies), after diluting whole blood 1:1 with PBS. The red blood cell and granulocyte fraction of the blood was then removed. The MNC fraction underwent red cell lysis using 1 incubation at 4°C for 15 mins with RBC lysis buffer (BioLegend). CD34 positive cell selection of peripheral blood and cord blood MNC samples was undertaken using the EasySep human whole blood CD34 positive selection kit (STEMCELL Technologies). The kit was used as per the manufacturer's instructions, but with only a single round of magnetic selection. Bone marrow MNCs did not undergo CD34 positive selection prior to cell sorting.

2.4 Fluorescence activated cell sorting

For all 'normal' samples MNC or CD34 enriched samples were centrifuged and resuspended in PBS/3%FBS containing an antibody panel consisting of: CD3/FITC, CD90/PE, CD49f/PECy5, CD38/PECy7, CD33/APC, CD19/A700, CD34/APCCy7, CD45RA/BV421 and Zombie/Aqua. Cells were stained (30 minutes at 4°C) in the dark before washing and resuspension in PBS/3%FBS for cell sorting. For all samples haematopoietic stem cell/multipotent progenitor (HSC/MPP) cells (Lin-, CD34+, CD38-, CD45RA-) were sorted using either a BD Aria III or BD Aria Fusion cell sorter (BD Biosciences) at the NIHR Cambridge BRC Cell Phenotyping hub. The gating strategy is illustrated in **Figure 2.1**. The 'HSC/MPP' population was treated as a single entity

and not further subclassified in the analysis. The immunophenotypic HSC/MPP population includes both long-term, intermediate-term and short-term HSCs as well as multipotent progenitors (MPPs), as demonstrated functionally in xenotransplantation assays[110–113].

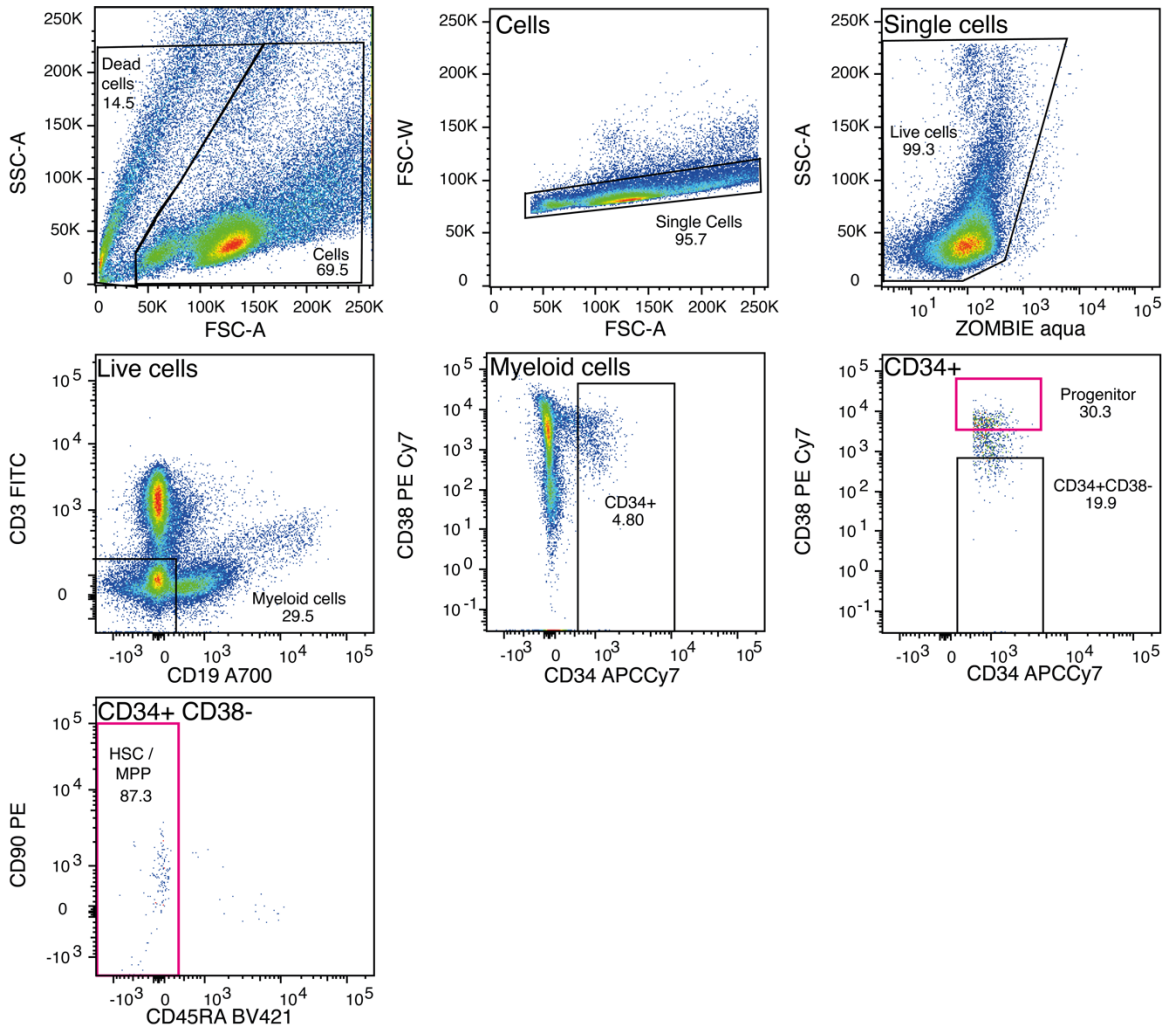


Fig. 2.1| Flow-sorting strategy for single HSC/MPP and HPC cells. a, Sorting of single human HSC/MPP and HPCs from cord blood, peripheral blood and bone marrow. Cells were stained with the panel of antibodies in **Table S1** then single HSC/MPP or HPCs were index sorted according to the strategy depicted into individual wells of 96 well plates.

In a subset of individuals, a small number of haematopoietic progenitor cells (HPCs; Lin-, CD34+, CD38+) were also sorted. The antibody panel used is shown in **Table S1**. The HPC cells were treated as a single entity in the analysis and not further subclassified. The

immunophenotypic HPC compartment includes predominantly myeloid and erythroid progenitors.

2.5 Single-cell colony expansion in liquid culture

For all 'normal' samples single phenotypic 'HSC/MPP' or 'HPC' cells were index sorted, as above, into Nunc 96 flat-bottomed TC plates (Thermofisher), containing 100μl supplemented StemPro media (Stem Cell Technologies) but no murine cell feeder layer. The following supplements were added to promote proliferation and push differentiation toward granulocyte, monocyte, erythroid and NK cell types: StemPro Nutrients (0.035%, Stem Cell Technologies), L-Glutamine (1%, ThermoFisher), Penicillin-Streptomycin (1%, ThermoFisher) and cytokines (SCF, 100 ng/ml; FLT3, 20 ng/ml; TPO, 100 ng/ml; EPO 3 ng/ml; IL-6, 50 ng/ml; IL-3, 10 ng/ml; IL-11, 50 ng/ml; GM-CSF, 20 ng/ml; IL-2 10 ng/ml; IL-7 20 ng/ml; lipids 50 ng/ml). Cells were incubated at 37°C and the colonies that formed were topped up with 50μl StemPro media plus supplements at 14 +/- 2 days as necessary. At 21 +/- 2 days a visual size assessment of colonies was undertaken prior to harvesting of cells for DNA extraction. Larger colonies (≥ approximately 3000 cells in size) were transferred to fresh U bottomed 96 well plate (Thermofisher). The U bottomed plates were then centrifuged (500 x g for 5 min), media was discarded, and the cells were resuspended in 50μl PBS prior to freezing at -80°C. Smaller colonies (<3000 but >200 cells in size) were harvested into 96 well skirted LoBind plates (Eppendorf) and centrifuged (800 x g for 5 min). Supernatant was removed to 5-10ul using an aspirator prior to DNA extraction on the fresh pellet. For larger colonies DNA extraction was performed using the DNeasy 96 blood and tissue plate kit (Qiagen). The Arcturus Picopure DNA Extraction kit (ThermoFisher) was used to extract DNA from the smaller colonies. Both kits were used as per the manufacturer's instructions. Overall, 42-89% of the sorted HSC/MPP population in each individual produced a colony (**Fig. 2.2a**). This is much more efficient than previously used methods of colony growth in semi-solid media, where it is has been estimated that 10% of single HSC/MPPs will produce a colony. In addition, analysis of cell surface markers showed there was no immunophenotypic difference between sorted HSC/MPPs that formed colonies and were sequenced, compared to those that did not (**Fig. 2.2b**).

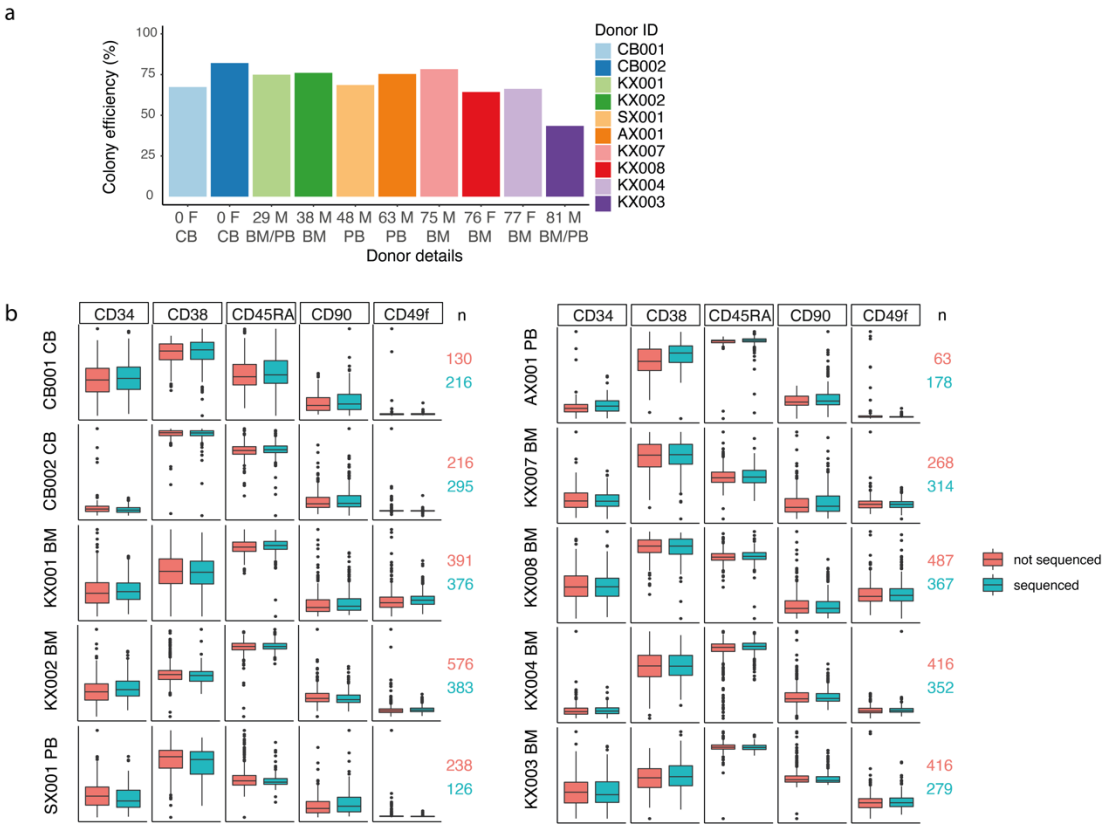


Fig. 2.2| Colony efficiency and cell surface marker analysis. a, Colony forming efficiency per individual of all single HSPCs sorted. **b,** Box-and-whisker plots showing fluorescence intensity for different cell surface markers used to define human HSCs, with different patients in rows. CD90 and CD49f are markers used to define the short and long term HSC subsets, which were included in our panel but were not used in sorting. Cells that produced colonies large enough to sequence are shown in teal; cells that did not form large enough colonies to sequence are in orange. The horizontal lines denote the median, the boxes the interquartile range and the whiskers the range.

2.6 Immunophenotyping of flow-sorted colonies

Following 21 +/- 2 days in culture, ½ of each colony selected by size criteria (> 3000 cells) was harvested into a U bottomed 96 well plate (ThermoFisher). Plates were then centrifuged (500g/5 minutes), media was discarded. After washing cells were resuspended in 100 µl PBS/3%FBS (30 min/4C) containing an antibody panel consisting of: CD45/PECy5, CD41/FIITC, CD11b/APCCy7, CD14/PECy7, CD15/BV421, GlyA/PE, CD56/APC (**Table S1**). Immunophenotyping of the mature cells in each stained colony was performed using a BD Fortessa2 (BD Biosciences) as per the gating strategy in **Figure 2.3a**. In the analysis ≥ 50 cells were required in a gate to call that cell type present. The one exception was the NK cell gate

for which a lower threshold of ≥ 30 cells were used. **Figure 2.3b** shows how the mature cell types produced in the assay are related to each other.

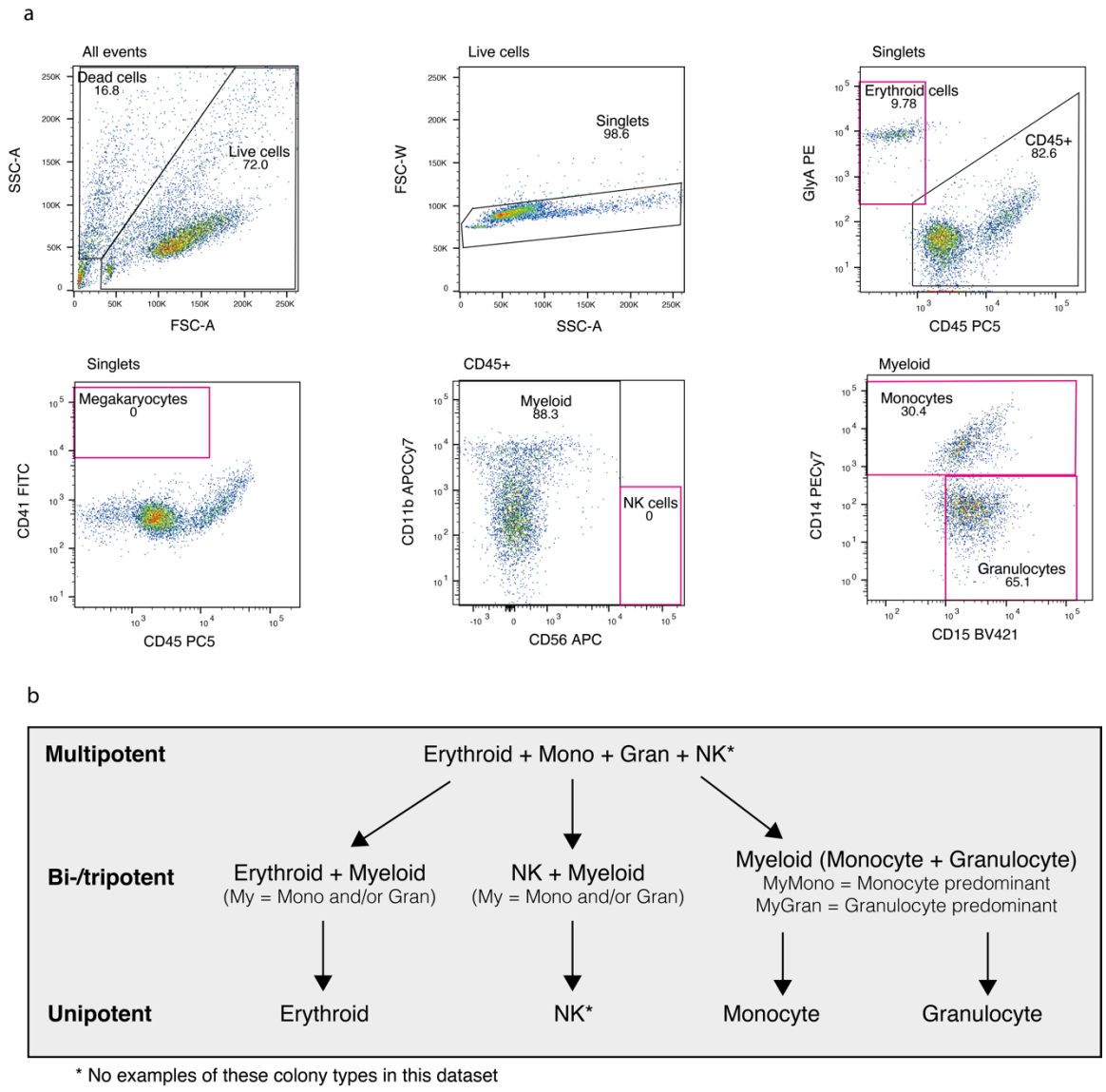


Fig. 2.3| Mature cell phenotyping of colonies. a, Gating strategy used for immunophenotyping of mature cells in single HSPC-derived colonies. The example colony is classified as EryMy (Erythroid + Mono + Gran). **b**, Classification scheme for the different colony phenotypes (note that multipotent colonies and NK-only colonies were not observed in this dataset).

2.7 Single-cell colony expansion in MethoCult

For all the chemotherapy-exposed individuals, MNCs were plated at a density of 7.5-45-x10^4/ml in MethoCult (H4435, STEMCELL Technologies) and incubated at 37C for 14 days. The cell suspensions were made up in StemSpan II (STEMCELL technologies) before being mixed thoroughly with MethoCult and plated into a SmartDish (STEMCELL technologies). Individual BFU-E or CFU-GM colonies were picked into 17ul of proteinase K (PicoPure DNA extraction kit,Fisher Scientific- each vial lyophilised proteinase k resuspended in 130ul reconstitution buffer) and incubated 65C for 6hrs, 75C for 30mins to extract DNA in preparation for sequencing.

2.8 Whole genome sequencing of colonies

A recently developed low input enzymatic fragmentation-based library preparation method[114,115] was used to generate whole genome sequencing libraries from 1-5ng extracted DNA from each colony. Whole genome sequencing for the 'ageing cohort' was performed at a mean sequencing coverage of 14X (8-35X) on either the Hiseq X or the NovaSeq platforms (Illumina). BWA *mem* was used to align 150bp paired end reads generated to the human reference genome (NCBI build37). Whole genome sequencing for samples in the 'chemotherapy cohort' was performed at a mean sequencing coverage of 23X for the chemotherapy exposed colonies (19 individuals, 151 samples) and 24X for the comparator normal colonies (11 individuals, 110 samples).

2.9 Single-base-substitution and indel calling

CaVEMan and Pindel

CaVEMan (used for calling single nucleotide variants - SNVs) and *Pindel* (used for calling small indels) were run against an unmatched synthetic normal genome using in-house pipelines[116,117]. *CaVEMan* was run with the 'normal contamination of tumour' set to 0.05, otherwise standard settings were used. Default filters were also used, one of which excludes putative SNVs that are present in a large panel of normal samples, so excluding most of the germline single nucleotide polymorphisms (SNPs) from subsequent analysis. It leaves around 30,000-40,000 germline SNPs in most individuals, which represent inherited SNPs that are rare within the population. In addition to the default *CaVEMan* filters, thresholds were set to

require putative variants to have a mean mapping score (ASMD) of at least 140 and fewer than half supporting reads being clipped (CLPM=0). *Pindel* was run with standard settings.

'Low input' filtering

A custom filter was then used to remove artefacts associated with the 'low input' library preparation method, including those due to cruciform DNA structures[57]. Specifically, the custom 'low input' filter incorporates two additional filtering strategies. Firstly, a fragment-based filter, designed to remove overlapping reads that result from the relatively shorter insert sizes produced by this protocol, which can result in the double counting of variants. Secondly, a cruciform filter, which removes erroneous variants introduced due to the incorrect processing of cruciform DNA. For each variant, the standard deviation (SD) and median absolute deviation (MAD) of the variant position within the read was calculated separately for positive and negative strands reads. Where a variant was supported by a low number of reads for one strand, the filtering used statistics calculated from the reads derived from the other strand. It was required that either: (a) ≤90% of supporting reads report the variant within the first 15% of the read as determined from the alignment start, or (b) MAD > 0 and SD > 4. Where both strands were supported by sufficient reads, it was required for both strands separately to either: (a) ≤90% of supporting reads report the variant within the first 15% of the read as determined from the alignment start, (b) MAD > 2 and SD > 2, or (c) at least one strand has MAD > 1 and SD > 10.

cgpVAF

Following this, *cgpVAF* (another bespoke algorithm) was used to generate a matrix of variant and normal reads at all sites that had a detected variant in any sample from a given individual. These algorithms are available from the Sanger Institute's Cancer IT GitHub repository (https://github.com/cancerit).

Variant filtering: large sample numbers per individual

Filtering on the read count and depth matrices containing 40 to several hundred samples per individual was then performed as follows: a) An exact binomial filter was used to remove variants with aggregated count distributions consistent with germline single nucleotide

polymorphisms (SNPs)[118]. b) A beta-binomial filter was used to remove low-frequency artefacts, i.e. variants present at low frequencies across samples in a way not consistent with the sample-to-sample variation expected for acquired somatic mutations[118]. c) Sites with a mean depth below 8 and over 40 were removed. d) Thresholds for read count and VAF were used to filter out *in vitro* variants from the remaining mutations using a bespoke script. The thresholds were set to require a minimum variant read count of 2 or more and a variant allele fraction of 0.2 for autosomes and 0.4 for XY chromosomes (**Fig. 2.4**). e) For each site normal and variant read counts were aggregated from samples with ≥ 3 variant reads. A one-sided exact binomial test was used to filter mutations inconsistent with a true somatic mutation (p-value < 0.001). f) A final filtering step was the removal of mutations that best mapped to the 'ancestral' branch of the SNV-derived phylogenetic tree (only the case for 8 mutations in one individual). Custom R scripts, used for these filtering steps were adapted from Spencer Chapman *et al*[119] (https://github.com/emily-mitchell/normal_haematopoiesis/2_variant_filtering_tree_building/scripts/).

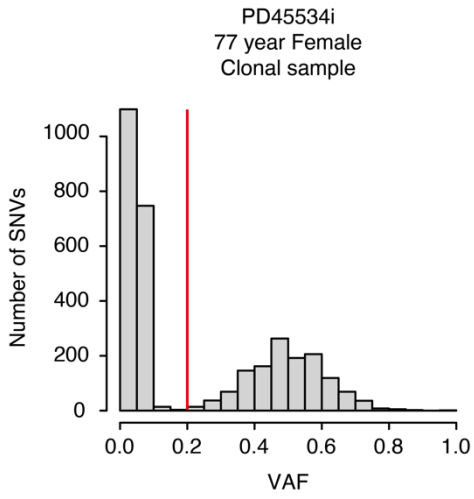


Fig. 2.4| Variant allele frequency threshold. Histogram of VAFs for a typical sample in the dataset, showing a tight distribution around 50%, as expected for an uncontaminated clonal sample derived from a single cell. The variants with VAFs < 0.2 represent *in vitro* acquired mutations and sequencing artefacts and were removed using a VAF-based filtering strategy with a cut off of 0.2 (red line).

Variant filtering: small sample number per individual (chemotherapy dataset only)

Additional filtering on the read count and depth matrices containing 4-10 samples per individual (sequenced at 23-24X depth) was performed as follows: a) Remove variants present in ≥ half colonies as likely germline. b) Call variants as present in a given sample if there are at least 2 supporting reads and the VAF is ≥ 0.2 for autosomes and ≥ 0.4 for XY chromosomes (same criteria used above). c) Sites with a mean depth below 8 and over 50 were removed. This adapted approach was required as the binomial filtering approach described above was not possible with such small numbers of samples per individual.

2.10 Structural variant and copy-number calling

Structural variant calling

Written by Hyunchul Jung

Structural variants (SVs) were called using GRIDSS[120] (Genome Rearrangement Identification Software Suite), with all variants confirmed by visual inspection and by checking if they fit the distribution expected based on the SNV-derived phylogenetic tree. Specifically, GRIDSS with a default setting (version 2.9.4) was used to call SVs. SVs larger than 1kb in size with QUAL >=250 were included. For SVs smaller than 30kb, SVs with QUAL >=300 were only included. Furthermore, SVs that had assemblies from both sides of the breakpoint were only considered if they were supported by at least four discordant and two split reads. SVs were further filtered out for which the standard deviation of the alignment positions at either ends of the discordant read pairs was smaller than five. To remove potential germline SVs and artefacts, a panel of normal was generated by adding in-house normal samples (n=350) to the GRIDSS panel of normal. SVs found in at least three different samples in the panel of normal were removed.

Copy number calling

Autosomal copy number aberrations (CNAs) and X chromosome CNAs in females were called using another in-house algorithm, ASCAT (Allele-Specific Copy number Analysis of Tumours)[121], which was run against a single sample selected from each individual. The matched sample was selected to have a coverage > 15X, no loss of Y and to be a singleton in the phylogenetic tree (no coalescences post birth). The ASCAT output was manually

interpreted through visual inspection. ASCAT was unable to accurately call copy number changes on the haploid sex chromosomes in males. Therefore, the in-house algorithm BRASS (BReakpoint AnalySiS)[122] was run to generate an intermediate file containing information on binned read counts across 500bp segments of the genome. A comparison of the mean coverage of the X and Y chromosomes was used to call X or Y CNAs in individual samples, which were then validated by visual inspection of read depth.

2.11 Filtering at the colony level

Ageing dataset

Some colonies were removed from the ageing dataset due to low coverage (17 samples), being technical duplicates (34 samples) and for showing evidence of non-clonality or contamination (7 samples). A peak VAF threshold of < 0.4 (after the removal of *in vitro* variants) was used, as well as visual inspection of the VAF distribution plots, to identify colonies with evidence of non-clonality (**Fig. 2.5**). Visual inspection was particularly important in the cord blood samples where there was greater variability in the distribution of variant allele frequencies due to the lower mutation burden. In these samples the VAF threshold of < 0.4 was therefore less stringently applied.

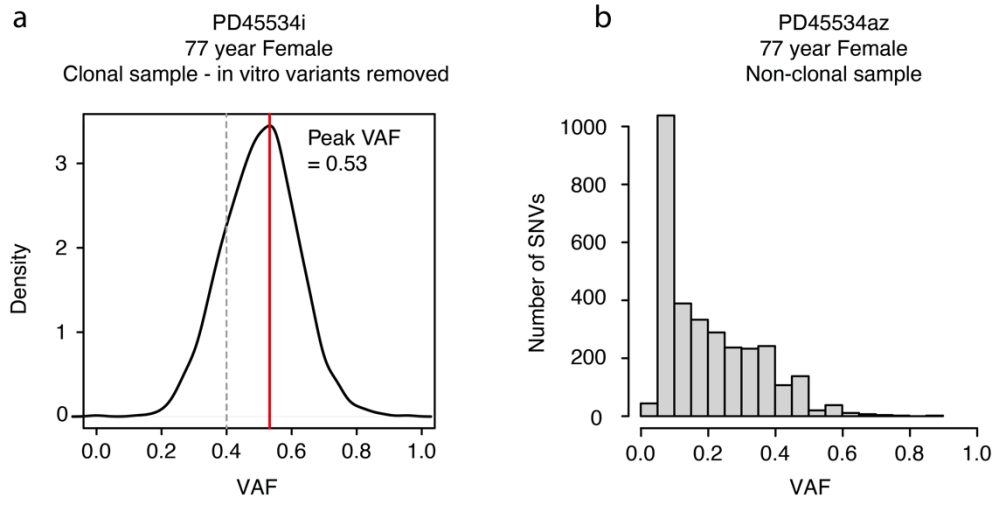


Fig. 2.5| Assessment of sample clonality. a, VAF distribution of variants after filtering steps had been applied. The red line shows the peak VAF and the dashed grey line shows the threshold peak VAF for excluding samples as being non-clonal / contaminated. **b,** Histogram of VAFs for a colony

that was seeded by 2 cells showing a median VAF around 25%. Colonies showing evidence of non-clonality in this way were excluded from downstream analysis using a peak VAF cut off of 0.4.

Chemotherapy dataset

The filtering strategies above were also applied to the 'mutation burden analysis' samples sequenced at higher coverage in the chemotherapy cohort. A total of 4 samples were removed from this smaller dataset due to evidence of non-clonality.

The 'population structure analysis' samples, also underwent the filtering strategies above. A total of 63 samples were removed from the 'population structure analysis' sample dataset due to being duplicates (22 samples), low coverage (16 samples) and non-clonal (25 samples). For the 'population structure analysis' samples, an additional filtering strategy based on the VAF of mutations across the phylogeny was also used. *This method was developed by Mike Spencer Chapman who wrote the following text:* Each sample was tested against the phylogeny to see if the mutation VAFs across the tree were as expected for a clonal sample. A clonal sample should have either branches that are 'positive' (mutation VAFs ~0.5), or 'negative' (mutations VAFs ~0). Therefore, for each branch in each sample, variant and total read counts were combined across all branch mutations. These counts were then tested for how likely they were to come from either (1) at least that expected for a heterozygous somatic mutation distribution, with some contamination allowed (one-sided exact binomial test, alternative hypothesis = less than probability, probability = 0.425), or (2) no more than that expected for absent mutations, with some false positives allowed (one-sided exact binomial test, alternative hypothesis = greater than probability, probability = 0.05). If samples had any branches with read counts that were highly inconsistent with both tests (maximum q-value <0.05, Bon-Ferroni correction) or had 3 or more branches that were minorly inconsistent with both tests (maximum p-value 0.05, no multiple hypothesis testing correction) the sample was considered non-clonal and excluded. A second iteration of phylogeny building was then performed without the non-clonal samples. This approach only identified 1 additional non-clonal sample included in the total above.

2.12 Validation of mutation calls

Mutation spectrums generated from the normal individuals were consistent with previously published data[3,53,123](**Fig. 2.6a**). Mutation spectrums were compared between the set of shared mutations (those present in 2 or more colonies), which are those we have the greatest confidence in, and private mutations (present in only one sample), in which we have lower confidence (**Fig. 2.6b**). The mutation spectrums are almost identical, providing evidence that the private mutation set does not contain excess artefacts.

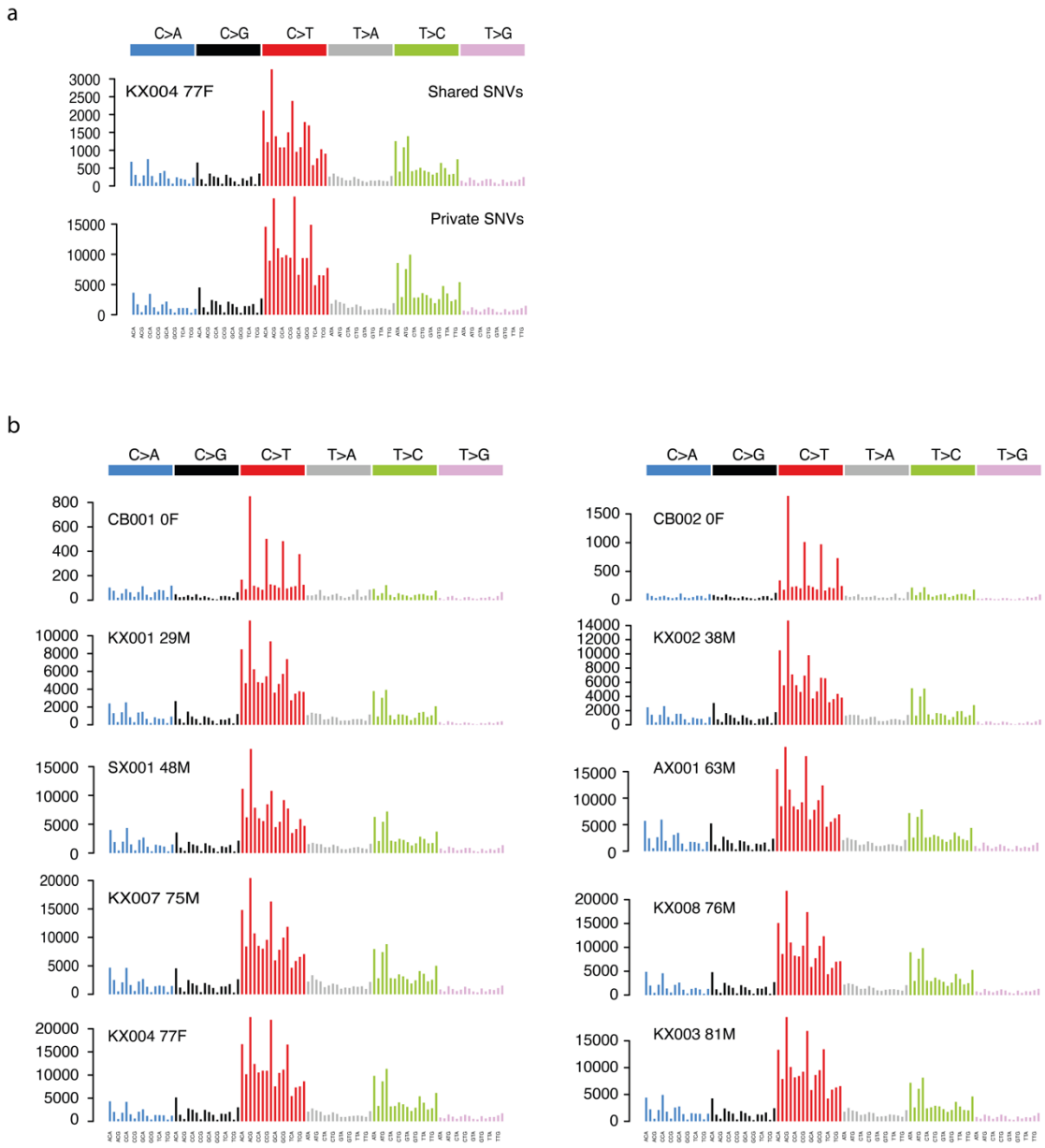


Fig. 2.6| Validation of mutation calls. a, Trinucleotide context mutation spectra of private (top plot) and shared variants (bottom plot) for one individual. The spectra are extremely similar, showing

the variant filtering strategy used is robust and prevents excess artefacts in the private variant set. **b,** Trinucleotide mutation spectrums for each individual created from all variants post filtering. The results are consistent between the two cord blood donors and all the adult donors.

2.13 Mutation burden analysis

Ageing dataset

SNV and indel burden analysis for the 'ageing' dataset was performed by first correcting the mutation and indel burden to a sequencing depth of 30, by fitting an asymptotic regression to the data (function *NLSstAsymptotic*, R package *stats*) (**Fig. 2.7**). *The code used to perform this was adapted from a script written by Heather Machado.*

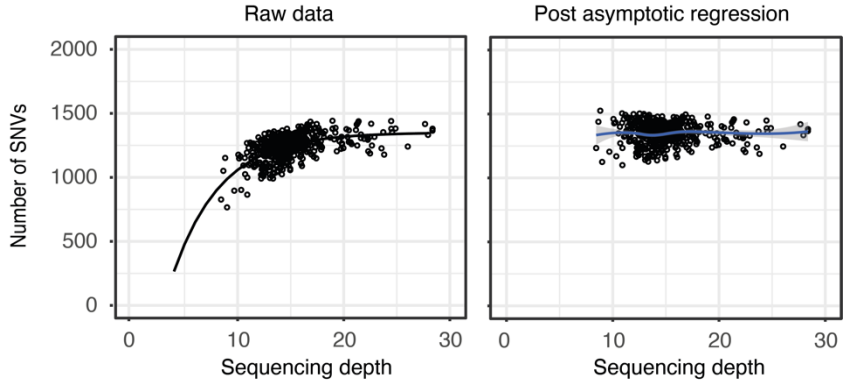


Fig. 2.7| Asymptotic regression to adjust for sequencing depth. Left-hand plot shows the relationship between raw mutation counts per colony for one individual post filtering and sequencing depth. The black line depicts an asymptotic regression line fitted to the raw data. Right-hand plot shows the adjusted mutation burdens per colony after asymptotic regression correction.

Subsequently, linear mixed effects models were used to test for a linear relationship between age and number of SNVs or number of indels (function *lmer*, R package *lme4*). Number of mutations or indels per colony was regressed using log-likelihood maximisation and age as a fixed effect, with the interaction between age and donor as a random effect. Progenitor samples were excluded from this analysis.

```
age.mut <- lmer(sub_adj ~ age + (age | donor_id), data = summ_cut[summ_cut$cell_type ==
"HSC",], REML = F)
```

```
age.indel <- lmer(indel_adj ~ age + (age | donor_id), data = summ_cut[summ_cut$cell_type == "HSC",], REML = F)
```

I also performed linear regression of age and mutation burden as above broken down by age range. Using data from the two cord blood donors and the youngest adult age 29 gives a rate of mutation accumulation of 17.56 per year ($CI_{95\%}$= 17.32-17.78). Using data from just the 3 younger donors aged 29, 38, 48 and 63 gives a rate of mutation accumulation of 17.21 per year ($CI_{95\%}$= 16.12-18.3). Linear regression of age and mutation burden using the 5 older donors aged 63, 75, 76, 77 and 81 gives a rate of mutation accumulation of 18.84 per ($CI_{95\%}$= 16.82-20.86). These results are very consistent over different phases of life.

Chemotherapy dataset

Due to the difficulty in correcting for sequencing depth when only a small number of samples is sequenced per individual, mutation burden analysis was performed on raw data (without asymptotic regression correction). This was possible due to the higher coverage of sequencing used for these samples (mean 23X, range 15X-33X). **Figure 2.7** shows that sequencing depth has little impact on mutation burden over this higher range. Linear mixed effects models were used to test for a linear relationship between age, chemotherapy for Non-Hodgkin's lymphoma (NHL) or colon cancer and number of SNVs (function *lmer*, R package *lme4*). Number of mutations per colony was regressed using log-likelihood maximisation and age and chemotherapy as a fixed effects, with the interaction between age and donor as a random effect.

```
NHL.sig <- lmer(Number_mutations ~ Age + NHL + (Age - 1|PDID), data = (subset(Summary_All, Summary_All$Cancer_diagnosis %in% c("Normal", "Follicular lymphoma", "DLBCL"))), REML = F)
```

```
Colon_cancer.sig <- lmer(Number_mutations ~ Age + Colon_cancer + (Age - 1|PDID), data = (subset(Summary_All, Summary_All$Cancer_diagnosis %in% c("Normal", "Colon cancer"))), REML = F)
```

2.14 Telomere analysis

Ageing dataset

Telomere length for each colony sequenced on the Hiseq platform (corresponding to the telomere length in the founding HSC/MPP or HPC) was estimated from the ratio of telomeric to sub-telomeric reads using the algorithm *Telomerecat*[124]. Colonies sequenced on the Novaseq platform could not be used as telomeric reads are removed by a QC step prior to bam file creation.

Linear mixed effects models were used to test for a linear relationship between age and telomere length across all the adults. Cord blood samples were excluded due to possible non-linearity of the relationship between age and telomere length in very early life[125].

```
age.tel <- lmer(tel_length ~ age + (age | donor_id), data = subset(summ_cut, summ_cut$platform == "hiseq" & !summ_cut$donor_id %in% c("CB001") & summ_cut$cell_type == "HSC"), REML = F)
```

Normality testing of the telomere length distributions was performed in R using the Shapiro-Wilk normality test (function *shapiro.test*) and visualised using Q-Q plots and density plots (functions *ggqqplot* and *ggdensity*). The percentage of outlying HSC/MPPs per individual was calculated using the interquartile range criterion (all samples outside the following interval are considered as outliers I = [q0.25−1.5·IQR; q0.75+1.5·IQR]) (function *boxplot.stats*). For all individuals the only outliers in the data had longer than expected rather than shorter than expected telomeres.

Chemotherapy dataset

Telomere analysis could not be performed on the chemotherapy exposed dataset, as no samples were sequenced on the Hiseq platform (all Novaseq).

2.15 Construction of phylogenetic trees

MPBoot, a maximum parsimony tree approximation method[126], was used to build phylogenetic trees of the relationships between the sampled cells. Variants were genotyped

as ‘present’ (coded as 1) in a sample if 2 or more variant reads supported the variant. Variants were genotyped as ‘absent’ (coded as 0) in a sample if 0 variant reads were present at a given site and depth at that site was 6 or more. Sites that did not fall into either of the above categories were marked as ‘unknown’ (coded as 0.5). In all cases only a small minority of sites (< 5%) were categorised as ‘unknown’ or ‘missing data’ as shown in the **Table 2.1**.

Table 2.1: Percentage variant sites ‘missing data’

Sample_ID	Genotype 0 ‘absent’	Genotype 0.5 ‘unknown’	Genotype 1 ‘present’	% sites ‘missing data’
KX001	3068118	42512	22049	1.38
KX002	2719616	28599	20085	1.05
SX001	4003266	127878	32218	3.17
AX001	5086607	173657	39577	3.39
KX007	9081044	48041	97580	0.52
KX008	10248698	78247	157878	0.75
KX004	21179417	197625	249761	0.92
KX003	8993397	107236	187015	1.17

The genotype matrix of shared variants was converted to a ‘DNA string’ for each sample with ‘W’ representing a ‘wildtype’ position, ‘V’ a ‘variant’ position and ‘?’ representing ‘unknown’. The DNA strings were then used as the input for *MPBoot*, which outputs unscaled trees with uninformative branch lengths (**Fig. 2.8a**). A ‘dummy sample’ (called “Ancestral”) was explicitly added in to the DNA strings that *MPBoot* used, with non-mutant genotypes across all sites i.e. representing the genotypes of the reference genome. After tree construction the ‘ancestral’ branch was dropped prior to downstream analyses. A maximum likelihood approach and the original count data was then used to assign each mutation in an individual’s dataset to a branch in their *MPBoot* generated phylogenetic tree

(https://github.com/NickWilliamsSanger/treemut). Tree edge lengths were then made proportional to the number of mutations assigned to the branch (**Fig. 2.8b**).

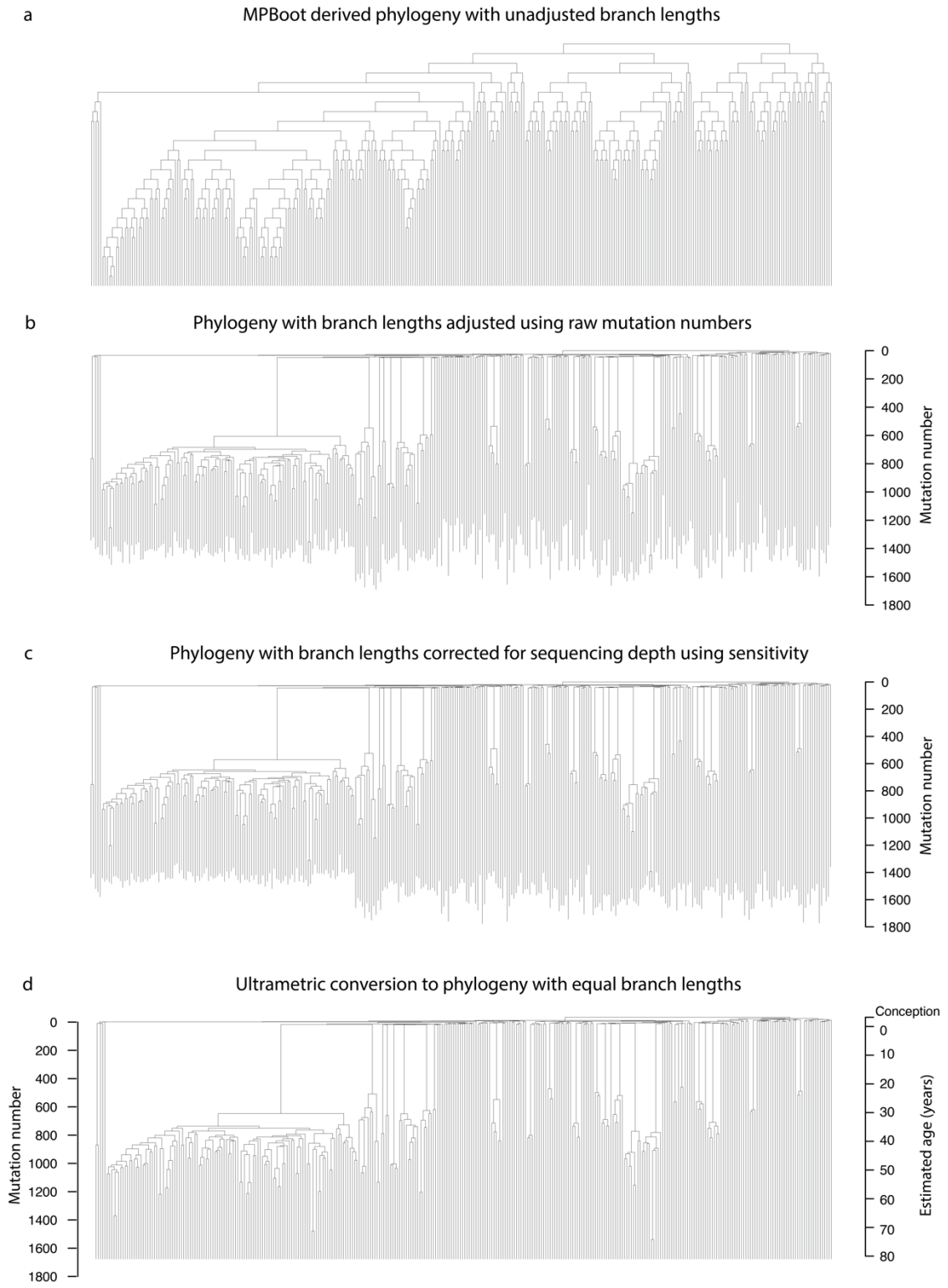


Fig. 2.8| Approach to phylogeny construction. a, Raw phylogeny for KX003 (81-year male) derived directly from *MPBoot*. The input to *MPBoot* is a genotype matrix of all variant calls shared by more

than 1 colony from an individual. **b,** Phylogeny with edge lengths proportional to the number of mutations assigned to the branch using original count data and the *tree_mut* package. **c,** Phylogeny with raw mutation count branch lengths adjusted for sequencing depth of the sample using sensitivity for germline variant calling. **d,** Phylogeny with adjusted branch lengths converted to ultrametric form (equal branch lengths). One axis shows mutation number, the other axis shows the equivalent estimated age in years, which is possible due to the linear accumulation of mutations in HSPCs with time. All tips end at age 81, the age at the time of sampling.

The sensitivity of mutation calling in each sample was used to correct phylogeny branch lengths for sequencing coverage. Sensitivity was calculated as the fraction of known germline variants identified by CaVEMan in a specific sample. Mutation burden was corrected by multiplying the number of variants by 1/sensitivity for private branches. The sensitivity was adjusted to allow for the higher sensitivity on shared branches due to multiple samples containing the variant. Specifically, sensitivity was assessed by measuring the ability of the mutation-calling algorithms to detect heterozygous germline single nucleotide polymorphisms (SNPs) in each sample. Heterozygous SNPs should have the same VAF distribution and sensitivity as true somatic mutations. For private branches, the SNV component of branch lengths was scaled according to:

$$= \frac{n_{SNV}}{p_i}$$

Where n_{cSNV} is the corrected number of SNVs in sample *i*, n_{SNV} is the uncorrected number of SNVs called in sample *i* and p_i is the proportion of germline SNPs called by the Caveman algorithm in sample *i*.

For shared branches, it was assumed that (1) the regions of low sensitivity were independent between samples, (2) if a somatic mutation was called in at least one sample within the clade, it would also be correctly called (or 'rescued') in other samples in the clade (even in lower sensitivity samples). Shared branches were therefore scaled according to:

$$\frac{n_{SNV}}{1 - \pi_i(1 - p_i)}$$

Where the product is taken for $1 - p_i$ for each sample *i* within the clade. However, both of these assumptions will not hold true in all cases. Firstly, regions with low coverage are not

randomly distributed, with some genomic regions likely to have low coverage in multiple samples. Secondly, while many mutations will be 'rescued' in subsequent samples once they have been called in a first sample - because the *treemut* algorithm for mutation assignment uses original read count data, meaning that even a single variant read in a subsequent sample is likely to result in the mutation being correctly assigned - this will not be true in every case. Some samples with very low coverage have 0 variant reads at a given site will by chance. In this situation, a mutation may not be correctly placed. While these factors may lead to an under-correction of shared branches, this approach provides a reasonable approximation. Corrected SNV burdens for each sample can then be calculated as the sum of corrected ancestral branch lengths back to the root of the phylogeny. **Figures 2.9 and 2.10** show the phylogenies with branch lengths corrected for differences in sequencing depth.

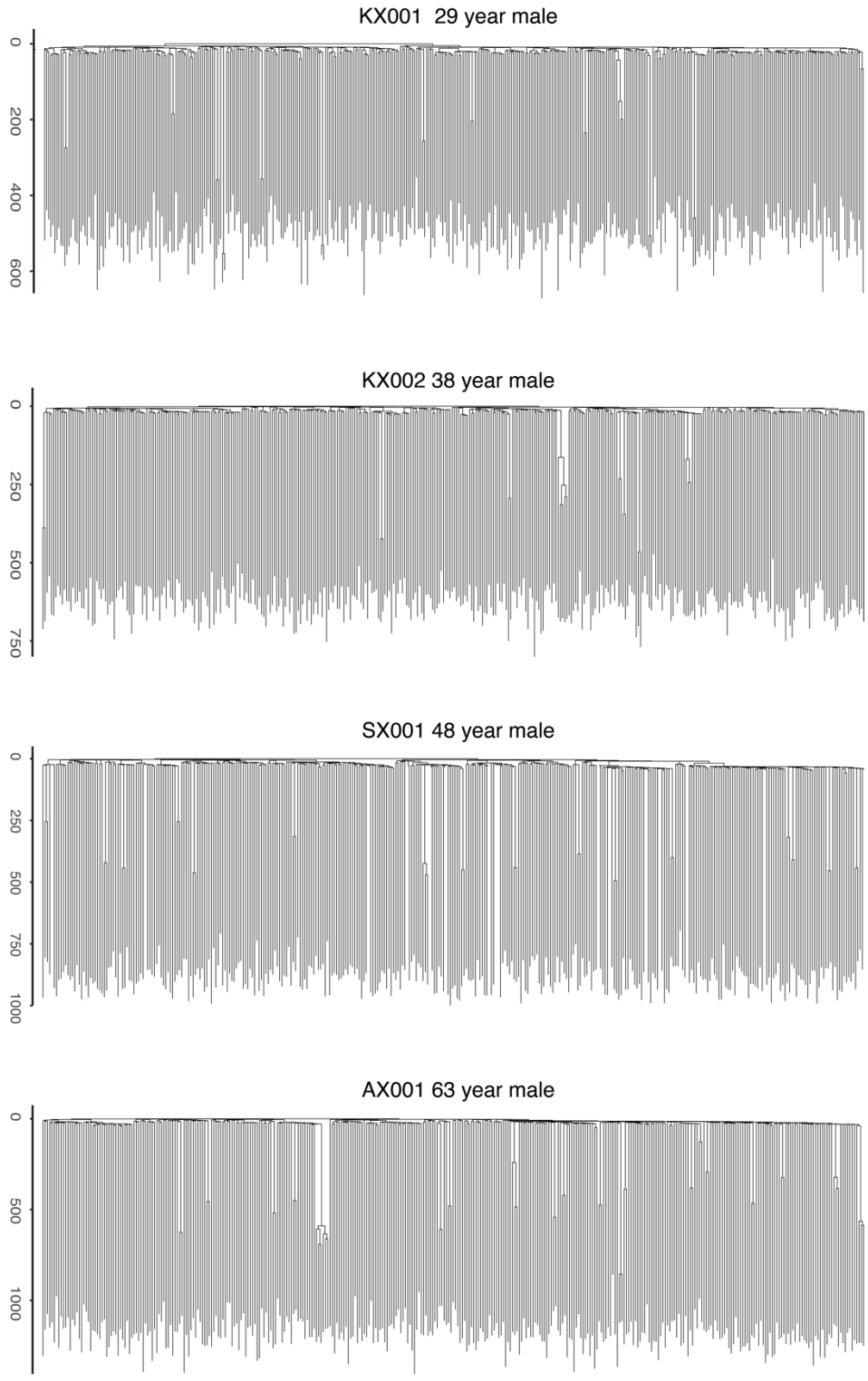


Fig. 2.9| Raw phylogenies for the four youngest adult donors. Phylogenies shown with raw mutation count branch lengths adjusted for sequencing depth of the sample using sensitivity.

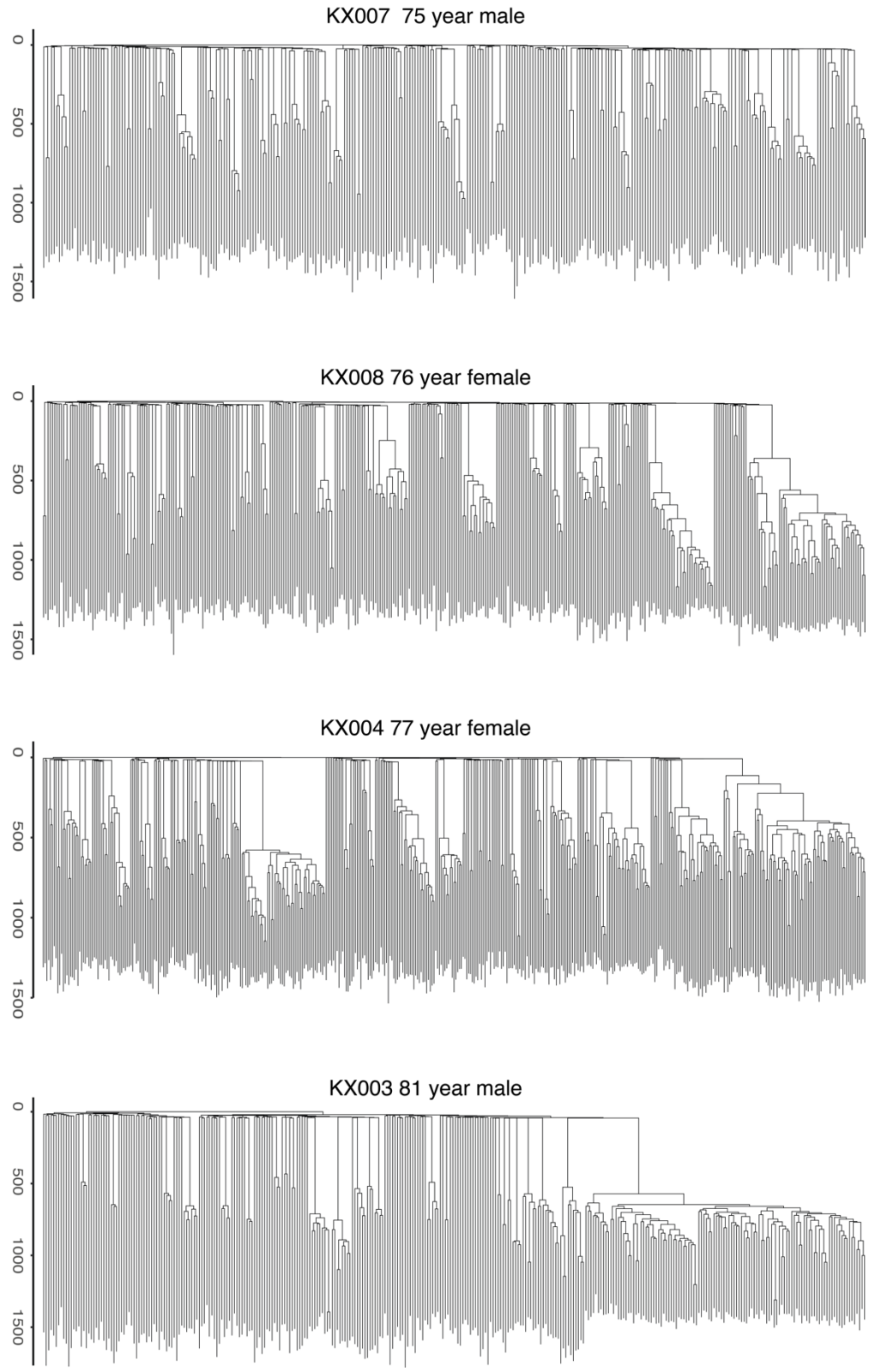


Fig. 2.10| Raw phylogenies for the four elderly adult donors. Phylogenies shown with raw mutation count branch lengths adjusted for sequencing depth of the sample using sensitivity.

The phylogenies were then made ultrametric (or linearised) using a bespoke algorithm to make all branch lengths equal (**Fig. 2.8c**, **Supplementary Code**). Starting from the root of the tree and moving progressively towards each tip, the fraction of time for the given shared branch is calculated as the fraction of remaining time times the number of mutations on the given shared branch divided by the mean number of mutations of all descendants from that shared branch. The function is called recursively, updating the fraction of remaining time, as the algorithm moves from root to tip. This algorithm therefore has the property that the most confident timings (nodes near the root) are defined first, anchoring the timings of subsequent, less confident nodes. It was desirable to convert the raw phylogenies (**Figs. 2.19 and 2.10**) into ultrametric phylogenies with equal branch lengths as this was a requirement for the *phylodyn* algorithm to function correctly. Ultrametric trees also provide the most reliable estimates of coalescent event timing, which was also important for the subsequent modelling work.

We compared the results obtained using our custom method for linearising the phylogenies and an alternative Bayesian approach (*Rtreefit*) utilised by Williams *et al*[127]. In brief, *Rtreefit* is a Bayesian model for converting mutation count based trees into time-based trees. The method jointly fits a global constant mutation acquisition rate and absolute time branch lengths under the assumption that the observed mutation count based branch lengths are Poisson distributed with $Mean = Duration \times Sensitivity \times Mutation\ Rate$ and subject to the constraint that the root to tip duration is the age at colony sampling. The mean branch timings are directly sampled from the posterior distribution and by construction the resulting trees are guaranteed to have a root to tip distance that matches the sampling age of the colony. The model is coded in R and Rstan and inferred using the Rstan implementation of Stan's No-U-Turn sampler variant of Hamiltonian Monte Carlo method. For each patient tree the model was fitted across four chains each with 20,000 iterations including 10,000 burn-in iterations. The code is available as an R package '*Rtreefit*' at https://github.com/NickWilliamsSanger/rtreefit.

We found extremely high concordance between our custom 'iteratively re-weighted means' approach and the Bayesian approach described above. In all phylogenies the R^2 for branch length comparisons between the two approaches was > 0.99 (**Fig. 2.11**).

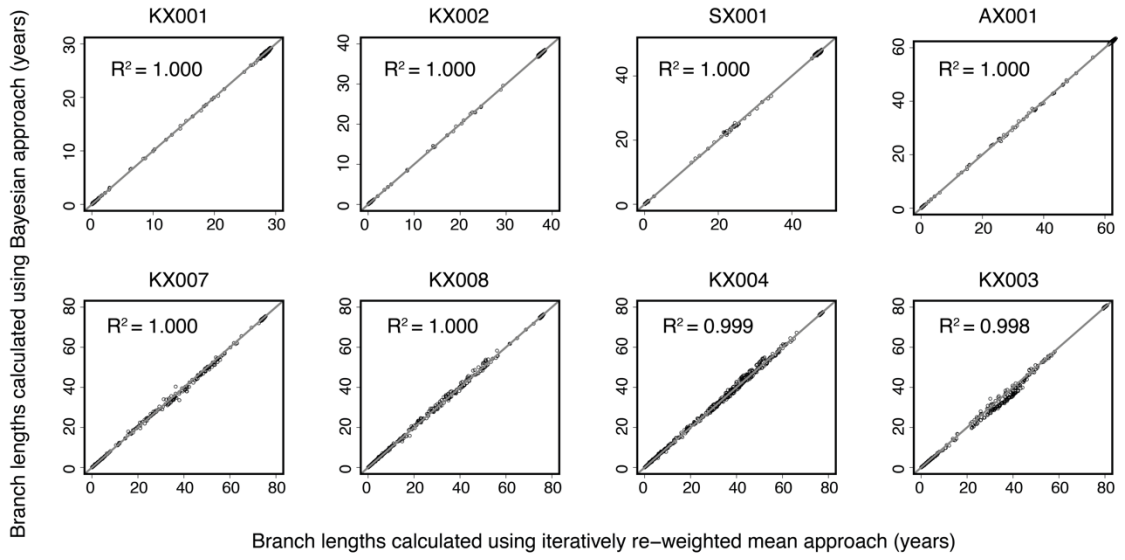


Fig.2.11| Comparison of phylogeny linearization methods. Plot comparing phylogeny branch lengths (in years) between the custom iteratively re-weighted means approach for phylogeny linearization used in this manuscript and an alternative Bayesian approach (Williams *et al*[128]).

Given the tight linear accumulation of mutations in HSPCs with age, the mutation branch lengths correspond to molecular time, which can be converted to time in years (**Fig. 2.8d**). Due to the known higher mutation rate during *in utero* development, which generates on average 55 somatic mutations in our cord blood HSC/MPPs, the first 55 mutations on the axis were assigned to the period between conception and birth (age 0), with the remaining mutation time evenly split between the years of age of the individual.

Additional information in the form of driver mutations, copy number changes and Y loss was then overlaid on the final ultrametric version of each phylogeny (as in **Fig. 2.8d**) to generate the final phylogenies depicted in **Figs. 3.8 and 3.9**. Driver mutations (which had already been assigned to a phylogenetic node using the tree_mut script above) were identified in the dataset by searching the VAGRENT annotations in the filtered_muts$COMB_mats.tree.build$mat matrix (in Rdata file annotated_mut_set_XXX). Copy number changes and Y loss events were identified as described in the relevant section above on a per sample basis, with this information read in to the tree_cut_analysis.Rmd script

in .csv format and subsequently used in phylogeny annotation as described in tree_cut_analysis.Rmd.
(https://github.com/emily-mitchell/normal_haematopoiesis/4_phylogeny_analysis/scripts/tree_cut_analysis.Rmd)

2.16 Validation of the phylogenies

Significant contributions by Michael Spencer Chapman

To assess the robustness, internal consistency and stability of the shared variants and inferred phylogenies we used three main approaches as outlined below.

Bootstrapping of the original read counts

One well-established approach to assessing the robustness of individual clades in a phylogeny is to repeatedly bootstrap the mutation matrix and re-build the phylogeny, observing in what proportion of bootstraps each clade is retained. *MPBoot* incorporates a bootstrap approximation method. However, somatic data has a well-established 'root' (the human reference genome) which makes this approach less applicable in our setting where we have high confidence in early splits that our supported by multiple samples, even if the numbers of mutations on the branch are low. With this data type the major cause of uncertainty is knowing exactly which cells carry a mutation, and what impact this would have on the inferred tree structure. Therefore, to better assess this type of uncertainty, an alternative bootstrapping approach as per Spencer Chapman *et al*[119] was used. Specifically, a partially-filtered mutation set was used to bootstrap the read counts for each colony at each locus. We then subjected the raw read count data to the same filtering and phylogeny-building approach as was used on the original data, with 1000 replicates per individual. The only exception was the beta-binomial filter. This was applied to the simulated data before the read count boot-strapping step. As with conventional approaches, the bootstrap phylogenies were then compared to the observed phylogeny to assess the proportion of bootstraps in which each clade is retained or lost. This was compared to the conventional mutation bootstrapping approximation performed by *MPBoot* (**Fig. 2.12a**). Quartet divergence and Robinson-Foulds similarities were calculated using the tqDist algorithm40 implemented in the R package Quartet v1.2.041. The bootstrapping analysis was performed for one of the elderly

adult HSPC phylogenies to ensure the finding of clonal expansions in the phylogeny was robust. The bootstrap phylogenies had high correlation to the observed phylogeny (**Fig. 2.12b**) with a median Robinson-Fould similarity of 0.951 and quartet divergence of 0.999.

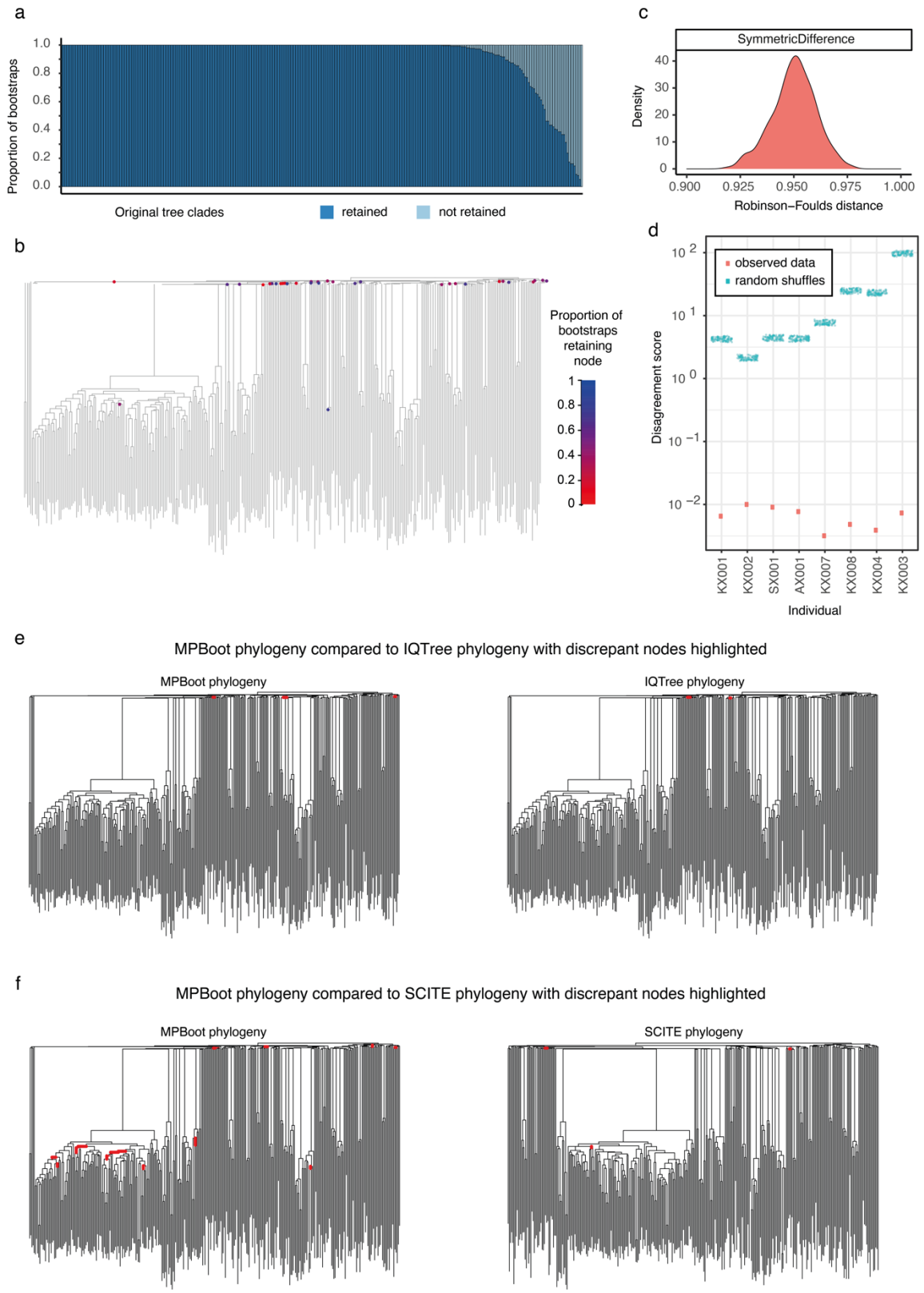


Fig.2.12| Phylogeny benchmarking. a, Robustness of each clade in the KX003 phylogeny (81-year male) using bootstrapping of the raw sequencing read count data. The proportion of bootstraps in which a clade is retained is shown, ordered by decreasing robustness. **b,** KX003 phylogeny annotated to show all nodes that have <90% bootstrap support using bootstrapping of the raw sequencing read count data. The nodes are highlighted with the average bootstrap support value.

c, Comparison of the sequencing read count bootstrap trees to the original trees by Robinson-Foulds similarity. **d,** Internal consistency of the genotype matrix for each adult individual as demonstrated by the disagreement score. The random shuffles have been displayed 'jittered'. A perfect phylogeny has a score of zero. **e-f,** Comparison of KX003 phylogenies generated by *MPBoot* and by the alternative phylogeny inference methods IQTree and SCITE. The nodes highlighted in red are those that differ between the inference methods, showing that the phylogenies are only minimally impacted by the inference method utilised. *Plots shown generated by Mike Spencer Chapman.*

Assessment of internal consistency of genotype matrices using the disagreement score

The disagreement score is based on the observation that in a perfect phylogeny any pair of mutations should either be in discrete clades or nested one within the other. To test the consistency of our data with this assumption a 'disagreement score' was used. For every pair of loci the number of cells in disagreement with this assumption was calculated. The mean score across all pairs was then calculated in such a way that cells with unknown genotypes were assumed to be in agreement. These scores from all the observed phylogenies were then compared to scores generated from random shuffles of the corresponding genotype table, internal to each locus. In this way the disagreement score in the observed genotype table can be compared to one that has been randomly generated. In all the observed trees the 'disagreement score' was extremely low compared to that obtained after random shuffling (**Fig. 2.12d**) showing our data has high internal consistency and the phylogenies are close to that expected in a perfect phylogeny (which would have a disagreement score of 0).

Comparison of the MPBoot phylogeny with other phylogeny inference methods

To assess the reliability and stability of the phylogeny generated by *MPBoot* we used the same genotype data as input into two alternative algorithms: IQ-TREE37 and SCITE38. IQ-TREE is a stochastic algorithm which infers phylogenies using maximum-likelihood. The Jukes-Cantor-type model for binary data was used as an appropriate model for single-cell whole-genome data. SCITE is an algorithm designed for somatic single-cell data. It uses Markov chain Monte Carlo sampling with an error model that takes potential false positives and false negatives into account for tree scoring. We used false positive and false negative rates of 0.001. Both these alternative algorithms produced phylogenies for KX003 (81 year male) with high agreement to the original *MPBoot* phylogeny with Robinson-Foulds distances of <0.05 for both comparisons (**Fig. 2.12e,f**).

To further explore whether using SCITE as an alternative phylogeny building approach would materially alter any conclusions, SCITE was run over all but one of the phylogenies. The largest phylogeny (KX004) could not complete within the timeframe required for the compute farm 'basement queue', which terminates jobs after 4 weeks. To be confident that the two methods give concordant trees, the similarity of trees estimated with MPBoot and SCITE was measured. Reassuringly, in all cases there was high concordance in the phylogenies produced by the two approaches (Robinsons-Foulds distance < 0.07) as shown in **Table 2.2**.

Table 2.2: Concordance of MPBoot vs SCITE phylogenies

Individual	Robinsons-Foulds Similarity of SCITE tree	Quartet Similarity of SCITE tree	Comparison of 32 summary statistics
KX001	0.934	1.000	Unchanged
KX002	0.949	0.999	Unchanged
SX001	0.945	0.999	Unchanged
AX001	0.947	1.000	Unchanged
KX007	0.977	0.999	Subtle changes (see below)
KX008	0.960	0.998	Unchanged
KX003	0.954	1.000	Subtle changes (see below)

Most differences that did emerge affected the precise arrangement of some early embryonic branch points – these differences would not be anticipated to have an impact on any of the key downstream analyses in the manuscript.

The summary statistics obtained from the phylogenies inferred using MPBoot vs SCITE were also formally compared. We found that when the range of summary statistics utilised for the driver ABC modelling were assessed at 4 timepoints, in 5 out of 7 individuals the statistics for the *MPBoot* and SCITE phylogenies were identical. For KX007, 6 of 32, and for KX003, 4 of 32 summary statistics calculated for each phylogeny are discordant but by a negligible amount (2 or less). These overall highly concordant findings further confirm that the choice of tree-

building approach would not have altered the conclusions of the downstream modelling analyses.

2.17 Inferring HSC population size trajectories

The R package *phylodyn*[90], provides a well-established approach to inferring historic population size trajectories from the pattern of coalescent events (more specifically the density of these events in historic time blocks) in a phylogenetic tree created from a random sample of individuals in the population. Its use has been pioneered in pathogen epidemiology[90] and has also been previously applied to HSPC data from a single individual[123]. **Figures 3.11 and 3.12** show how *phylodyn* can accurately recover simulated population trajectories using sample sizes similar to those we have used and illustrates how the number of coalescent events in a given time window in the tree informs on population size through time (assuming a constant rate of HSC symmetric cell division and a neutrally evolving population). We used *phylodyn* to infer historic changes in LT-HSC $N\tau$ from the ultrametric phylogenies of the four youngest adults in the cohort (**Fig. 3.10**).

2.18 Using rsimpop to simulate HSC populations

Written by Nick Williams

Simulations of complete HSC populations from conception to the age of sampling were performed for each individual using the R package *rsimpop*[127] (https://github.com/NickWilliamsSanger/rsimpop*). Rsimpop* utilises a birth-death model with specified somatic mutation accumulation rate and symmetric cell division rate, to simulate a complete HSC population. Each cell within the population has a rate of symmetric division and a rate of symmetric differentiation (or death). Asymmetric divisions do not impact on the HSC phylogeny and are not accounted for in the model.

Let α be the background rate of symmetric self-renewal cell divisions, measured in divisions per day. We model selective advantage of driver containing clone i as s_i. The increased rate of symmetric division $\alpha_i = \alpha(1 + s_i)$. We assume during the early population growth phase that the total population grows unrestrained by death. Once the specified population size, N,

is reached (within the first few years of life) then the death rate, β, for each cell matches the average division rate in the full population:

$$\sum^{cells} \beta = \sum^{\text{wild type cells}} \alpha + \sum_i \sum^{\text{cells in clone } i} (1 + s_i)\alpha$$

Thus giving

$$\beta = \frac{(N - \sum_i \mathrm{N}_i)\alpha + \sum_i \mathrm{N}_i(1 + s_i)\alpha}{N}$$

In the case of a single driver mutation containing clone with selection coefficient s then the deterministic phase behaviour is governed by a logistic growth function:

$$N_m = N \frac{1}{1 + exp\big(-\alpha s(t - t_m)\big)}$$

For some constant t_m (see Williams *et al)*[129]*.*

In the early stages of the exponential growth process, it exhibits an annual rate of growth S:

$S = exp(\alpha s) - 1.$

For multiple competing driver mutation containing clones, each with modest population sizes, it is expected that the above single clone approximation will apply for the individual competing clones. Once one or more of the competing clones represents a significant fraction of the overall population then the dynamics will be more complex. For cells containing more than one driver mutation the fitness effect on *S* is additive.

The above model is implemented using the Gillespie algorithm. The waiting time until the next event is exponentially distributed, with a rate given by the total division rate + total death rate. This event is then 'division' with probability=total division rate/(total division rate + total death rate). If the event is 'division' then the choice of which cell divides is given by a

probability proportional to the cell's division rate, whereas if the event is 'death' then all cells are equally likely to be chosen.

Implementation was in C++ with an R based wrapper as an R package *rsimpop*. The simulator maintains a genealogy of the extant cells, together with a record of the number of symmetric divisions on each branch, the absolute timing of any acquired drivers and the absolute timings of branch start and end. The package also provides mechanisms for sub-setting simulated genealogies whilst preserving the above per branch information.

2.19 HSC population size modelling

Significant contributions from Kevin Dawson

Simple neutral models of HSC populations (from which selection is absent) were investigated initially. The cell phylogenies, constructed from single cell genomes, include estimated branch lengths, from which we can calculate node heights, and hence the time intervals between successive coalescent events. In the case of a neutral model, the genomic data provides information about the trajectory of the product $N\tau$ (population size x time between symmetric self-renewal cell divisions).

However, the genomic data cannot provide information separately about N (population size), or τ (time between symmetric cell divisions). Furthermore, in the case of a neutral model, all the information provided by the genomic data, about the trajectory of the product $N\tau$, is contained in sequence of inter-coalescent intervals calculated from the phylogeny. This sequence of inter-coalescent intervals is precisely the information which the *phylodyn* package uses to infer the trajectory of the product $N\tau$.

Here, the aim was to perform additional Bayesian inferences about the parameters of neutral models from the phylogenies. Specifically, computation of marginal posterior densities (providing point estimates accompanied by credible intervals) was needed for the 'LT-HSC $N\tau$' parameter for the first 2-3 decades of life, and two additional parameters representing the midlife fold-change in $N\tau$ (elderly donors only), and late-life fold-change in $N\tau$ (all donors).

Flat prior densities on wide intervals were chosen (**Fig. 4.2**) to represent prior uncertainty about the values of these parameters, so that the resulting the marginal posterior densities could be compared with the inferences from the *phylodyn* package.

An additional motivation for performing these Bayesian inferences on neutral models, was to enable posterior predictive checks (PPC), in order to decide if the observed phylogenies are compatible with neutral models. Note that a separate donor-specific posterior distribution was generated (sampled) for each donor (donor-specific ABC), and a separate donor-specific posterior predictive p-value was computed for each donor (donor-specific PPC). Each donor-specific ABC for the neutral model was performed using the ABC rejection method (R package *abc*)[130,131].

We used the population trajectory from *phylodyn* to identify the time period prior to the increase related to a ST-HSC/MPP contribution, and the timing of the midlife and late-life fold-change in $N\tau$ (**Figs. 3.10 and 3.12**). This data was used to inform the choices for the time between symmetric cell divisions, which was set at 1 year (after the initial population growth phase in the first few years of life). The rate of mutation accumulation was set at 15 mutations per year with an additional 1 mutation for every cell division (both of these were drawn from a Poisson distribution centred on the input value).

In the younger individuals (aged < 65) estimates of $N\tau$ in the first few decades of life could be made due to the absence of the effect of positive selection (**Fig. 4.12**). However, in the older individuals (aged > 75), estimates of $N\tau$ could not be reliably calculated in the phylogenies due to the confounding effect of positive selection. Here the focus was on using the PPC method to decide whether the neutral model changes in population size (in the form of a bottleneck in the population in mid-life) is compatible with each of the observed trees.

The Bayesian inferences about the parameters of these neutral models were performed using Approximate Bayesian Computation (ABC) methods (in which large numbers of simulations of the data are performed using *rsimpop*, in place of computation of the likelihood function).

In order to apply these methods, the sequence of inter-coalescent intervals was replaced by a set of summary statistics (the 'number of lineages' in the tree through time at three points). For each donor, the marginal posterior densities for the parameters of interest are plotted alongside the corresponding prior densities, to illustrate how the data has reduced the uncertainty about the values of these parameters. For each donor, the sample from the (approximate) posterior distribution (generated by donor-specific ABC) was also used for the parameters of the neutral model, to perform donor-specific PPC.

A large sample of simulated data sets from the posterior predictive distribution was first generated, and from this a donor-specific posterior predictive p-value was estimated. The purpose of this donor-specific PPC is to decide if the observed phylogeny obtained from each donor is compatible with the proposed neutral model (while taking account of our uncertainty about the parameter values in the model). Here all features of the observed phylogenies are important (not only those features which are informative about the parameters of the neutral model). For observed phylogenies and simulated phylogenies, a chi-squared discrepancy variable can be calculated which incorporates many summary statistics (including clade size statistics). The posterior predictive p-value is computed from the upper tail-area probability under the distribution of the difference between the simulated chi-squared discrepancy and the observed chi-squared discrepancy[132]. If the p-value is close to zero, then the observed data is extreme (an outlier) compared to the data predicted under the proposed model (taking account of our uncertainty about the parameter values in the model). Thus, when the p-value is close to zero, this is evidence that the observed phylogeny is not compatible with the neutral model.

2.20 HSC population size estimate

A Monte Carlo simulation approach was used to sample from the distributions of each variable 500,000 times, calculating the value of N for each set of randomly sampled variables. This was done using the following distributions: telomeric shortening rate per division: uniform(minimum = 30, maximum = 100); symmetric division rate: uniform(minimum = 0.8,

maximum = 1.0); $N\tau$: uniform(minimum = 50,000, maximum = 250,000); average telomere loss per year: uniform(minimum = 30, maximum = 40).
(https://github.com/emily-mitchell/normal_haematopoiesis/6_population_modelling/scripts/estimating_N.Rmd)

2.21 Analysis of driver variants

Variants identified were annotated with VAGrENT (Variation Annotation GENeraTor) (https://github.com/cancerit/VAGrENT) to identify protein coding mutations and putative driver mutations in each dataset. **Table 3.2** lists the 17 genes we have used as our top clonal haematopoiesis genes (those identified by Fabre *et al*[133] as being under positive selection in a targeted sequencing dataset of 385 older individuals), whose 'oncogenic' and 'possible oncogenic' mutations (as assessed independently by myself and Peter Campbell) are shown in **Figures 3.8 and 3.9**. A more extensive list of 92 clonal haematopoiesis genes was also interrogated (**Table S2**)**.** In order to explore a wider set of cancer gene mutations we used the 723 genes listed in Cosmic's cancer gene census (https://cancer.sanger.ac.uk/census).

2.22 dN/dS analysis

Significant contributions from Peter Campbell

We used the R package *dndscv*[134] (https://github.com/im3sanger/dndscv) to look for evidence of positive selection in our dataset (https://github.com/emily-mitchell/normal_haematopoiesis/5_dNdS/scripts/all_DNDScv_final.Rmd). The *dndscv* package compares the observed ratio of missense, truncating and nonsense to synonymous mutations, with that expected under a neutral model. It incorporates information on the background mutation rate of each gene and uses trinucleotide-context substitution matrices. The approach provides a global estimate of selection in the coding variant dataset (**Table S3**), from which the number of excess protein coding, or 'driver mutations' can be estimated. In addition, it identifies specific genes that are under significant positive selection.

While a small bias in the estimated dN/dS ratio could lead to an apparently significant excess when the dataset contains large numbers of mutations, as our does. In defence of the

significant excess of non-synonymous mutations observed, there are 3 lines of supporting evidence:

1. Correction for confounders in the dN/dS algorithm

The dN/dS algorithm[135] is one of the best-in-class algorithms for quantifying somatic selection, as demonstrated by a recent pan-cancer comparison of different methods[136]. One of the reasons for this is the rigorous approach it takes to correcting for the known variables influencing mutation rate across the genome, including replication timing, chromatin state and DNAse accessibility[137]. In addition, the model balances the predicted mutation rates from these global covariates with the observed synonymous mutation rate within a gene – this latter correction captures many of the unknown variables affecting mutation rates acting at a local level.

Furthermore, the algorithm corrects for the observed mutational spectrum[135] – this is important because, for example, transitions are more likely to generate a synonymous mutation than transversions. The model parameterises all 192 rates representing the 6 different types of base substitution, the 16 combinations of bases 3' and 5' to the mutated base, and transcribed versus non-transcribed gene. This means that trinucleotide mutational signatures do not bias the overall dN/dS estimate.

2. Running dN/dS algorithm with greater stringency

In addition to running the dN/dS algorithm in its standard implementation, it was also run using two adaptations to impose greater stringency.

The first adaptation was to run the algorithm excluding sites that are masked by our variant caller in both the numerator and the denominator (a total of 175 million sites genome-wide). Essentially, most somatic mutation callers, including ours[116], have a 'normal panel' or equivalent where sites that are frequently non-reference because of sequencing artefact or germline polymorphism are masked. Since germline polymorphisms have a dN/dS ratio << 1, this can lead to under-calling of synonymous somatic mutations relative to non-synonymous mutations. Running the algorithm with sites in this normal panel excluded from both

numerator and denominator had minimal impact on the estimated overall value of dN/dS (1.0548, $CI_{95\%}$=1.02488-1.0856; versus 1.0586, $CI_{95\%}$=1.02861-1.0895 for the standard implementation). This argues that there is no bias arising from masking of true somatic mutations at germline polymorphisms.

The second adaptation was to run the dN/dS algorithm using correction for pentanucleotide sequence context. While a trinucleotide context captures virtually all of the effects of mutational signatures[138], there remains the theoretical possibility that any signature extending beyond that may affect synonymous mutations differently to non-synonymous mutations. To test this, the analysis was repeated using rates for the 6 mutation classes and 256 different combinations of 2 bases each side of the mutated base. This also had minimal impact on the estimated value of dN/dS for missense variants (1.0472, $CI_{95\%}$=1.0155-1.0799; versus 1.0589, $CI_{95\%}$=1.02852-1.0902 for the standard implementation) or the dN/dS for truncating variants (1.0788, $CI_{95\%}$=1.0106-1.1516; versus 1.0569, $CI_{95\%}$=0.99558-1.1220 for the standard implementation). Importantly, a pentanucleotide context covers the whole of the codon, no matter which base in the codon is mutated (whereas a trinucleotide context only covers the whole codon if the middle base is mutated) – this means that even if there were residual effects of mutational signatures beyond the pentanucleotide, they would not affect the mutated codon.

3. Measuring dN/dS on simulated mutations

As a further check, simulated mutations were generated in the sequencing data and the dN/dS algorithm was applied to them. 19 BAM files from cord blood HSC/MPPs in our dataset for which zero coding mutations were identified by our variant caller were selected. For each BAM file, 2000 sites in the exome were randomly chosen to have simulated mutations, with the mutations following the same mutational spectrum as observed in the whole dataset. At each position with a mutation, the reads reporting that base were extracted, and the base-call change recorded at that base with 0.5 probability (to get average VAF of 50%), according to the following rules: change to mutant base if read reported reference base; change to reference base if read reported mutant base; change to the other non-reference, non-mutant base if read reported non-reference, non-mutant base.

The modified BAM files then underwent exactly the same process of variant calling as our real data. The majority of the simulated mutations were correctly called (the proportion dependent on sequencing coverage), and the mutation spectrum was the same as that observed in the real data. In total, 29008 simulated mutations were called, which was very close to our real dataset of 25,888 coding mutations. The dN/dS algorithm was run over the simulated dataset and no bias was found in the results, with a dN/dS ratio for all randomly simulated variants of 1.00. For the simulated missense mutations, the estimated dN/dS was 1.001 ($CI_{95\%}$=0.974-1.028); and for simulated truncating mutations, it was 1.001 (0.956-1.067).

These simulations would have captured any biases in the estimation of dN/dS that arose from, for example, differential sequencing coverage across the genome, variant calling, variant filtering, variant annotation or the dN/dS algorithm. Instead, the estimates of dN/dS are almost exactly 1.00, as expected, with confidence intervals that do not overlap with those for our real data.

2.23 Amino acid variant annotation

Amino acid variant annotation was performed using SIFT4G (https://sift.bii.a-star.edu.sg/sift4g/AnnotateVariants.html)[139] and Polyphen2 (http://genetics.bwh.harvard.edu/pph2/bgi.shtml)[140]. Of a total 16536 missense mutations in our dataset, 5088 could be annotated by SIFT4G (38 in myeloid driver genes) and 4551 could be annotated by Polyphen2 (35 in myeloid driver genes). Approximately 42% and 45% of the annotated mutations were deemed to be 'deleterious' respectively (**Table 2.3**). If the same proportion of missense mutations is present in the dataset as a whole we would predict approximately 7000 'deleterious mutations', equating to around 1000 per adult individual or 2-3 per HSC.

Table 2.3: Amino acid variant annotation

	SIFT4G	Polyphen2
Total missense mutations in dataset	16536	16536
Number annotated	5088	4551

Number in known driver (excluded)	38	35
Deleterious	1998	2049
Possibly deleterious	319	762
Tolerated	2495	1709
Fraction deleterious	0.42	0.45
Predicted deleterious in whole dataset	6945	7441

Table S3 lists all coding variants used to run dN/dS with annotation including the SIFT and Polyphen2 scores.

2.24 Driver mutation acquisition rate estimation

Significant contributions from Peter Campbell

The dN/dS parameter is widely used in evolutionary genetics to infer patterns of selection[141,142], and has been recently adapted for cancer and somatic mutations[135]. It is essentially a measure of how far the observed number of non-synonymous mutations diverges from the number that would be expected from the synonymous mutation rate, after correction for mutation spectrum[135]. It is underpinned by the assumption that synonymous mutations evolve neutrally, and selection only acts on non-synonymous mutations. For example, a dN/dS ratio of 1 means that we observed exactly the same number of non-synonymous mutations as we would have expected for the number of synonymous variants. A dN/dS ratio of 2 means twice as many non-synonymous mutations as expected were observed, implying that half of the observed non-synonymous mutations occurred as expected for the background mutational processes, while the other half have accumulated through positive selection. From this, with a total number of observed non-synonymous mutations, the number of driver mutations in excess of the background expectation can be calculated (noting that this is an underestimate of the true number in the presence of any negative selection).

This is the intuition for the formal mathematical exposition. Given an observed number of non-synonymous mutations, n_{NS}, and an estimated dN/dS ratio, ω_{NS}, the formula for the expected number of drivers, n_{D}, is as follows:

$$n_{\mathrm{D}} = \frac{(\omega_{\mathrm{NS}} - 1)}{\omega_{\mathrm{NS}}} n_{\mathrm{NS}}$$

To give a worked example using missense substitutions in the ageing dataset, the overall dN/dS ratio was calculated to be 1.06 with a 95% confidence interval of 1.03 – 1.09 (**Figure 4.7a**). A total of 16,536 non-synonymous mutations was observed. The number of excess missense mutations can then be calculated as (1.06 – 1)/1.06 * 16536, which works out at 936, with the lower bound on the confidence interval as (1.03 – 1)/1.03 * 16536, which equals 482.

Linear mixed effects models were used to test for a linear relationship between age and the number of non-synonymous mutations. Colonies with a sequencing depth <14X were excluded.

age.non_syn.depth <- lmer(number_non_syn ~ age + (age | donor_id), data = subset(summ_cut, mean_depth > 14), REML = F)

This linear regression analysis found that non-synonymous mutations are acquired at a rate of 0.12/HSC/year ($CI_{95\%}$=0.11-0.13) and the dN/dS estimates inform that 1 in 12 to 1 in 34 non-synonymous mutations in the dataset are drivers. These estimates were used in a Monte Carlo simulation approach, sampling from the distributions of each variable 500,000 times, calculating the value of N for each set of randomly sampled variables. This was done using the following distributions: non-synonymous mutation acquisition per year: uniform(minimum = 0.11, maximum = 0.13); fraction of drivers: uniform(minimum = 0.029 (1/34), maximum = 0.083 (1/12)).
(https://github.com/emily-mitchell/normal_haematopoiesis/5_dNdS/scripts/estimating_driver_acquisition_rate.Rmd)

2.25 Y loss analysis

Significant contributions from Peter Campbell

The ageing dataset includes a series of phylogenetic trees from male individuals in which some clades have lost the Y chromosome. By eye, these clades seem to be larger than clades

that have not lost Y. To test this formally, a randomisation / Monte Carlo test was used to define the null expected distribution of clade size. For each Monte Carlo iteration, branches of the phylogenetic tree were drawn at random - one random branch for each observed instance of Y-loss. These branches were sampled (with replacement) from the set of all extant branches at the matched time-point in that individual, and the eventual clade size of that draw measured. For each simulation, the geometric mean (to allow for the log-normality of observed clade sizes) of clade sizes is calculated. The distribution of geometric means from the Monte Carlo draws can then be compared with the observed geometric mean. (https://github.com/emily-mitchell/normal_haematopoiesis/11_LOY_simulations/scripts/Loss_of_Y_simulations.Rmd)

2.26 Modelling positive selection in the HSC population

Significant contributions from Kevin Dawson

Given the evidence from the dN/dS analysis, the next task was to investigate more elaborate models of HSC population dynamics, incorporating positive selection acting on driver mutations. Here, as before, ABC methods were used to make inferences about the parameters of the model (incorporating positive selection), and posterior predictive checks (PPC), in order to decide if the observed phylogenies are compatible with this relatively simple non-neutral model (incorporating positive selection). In this non-neutral model, a static HSC population of 100,000 cells undergoing 1 symmetric self-renewal division per year, we explored a range of parameter values for the number of drivers introduced into the population per year, as well as the shape and rate of the gamma distribution used to define the distribution of fitness effects these drivers were drawn from (**Fig. 4.10**).

A threshold of 5% for the minimum fitness effect of these drivers was used (equivalent to a selection coefficient of 0.05) as Watson *et al* predicted drivers with a fitness effect of 4% or less could not expand to a VAF > 1% over the human lifespan[143]. We chose flat prior densities on wide intervals (**Fig. 4.10**) to represent prior uncertainty about the values of these parameters.

First, a separate donor-specific posterior distribution was generated (sampled) for each donor (donor-specific ABC). The simulations were performed using *rsimpop*, and the donor-specific ABC (ridge regression on the re-scaled, and logit-transformed, parameter values) was performed using the R package *abc*.

Second, a sequence of four ABC regression steps was used to generate a sample from the (approximate) multiple-donor posterior distribution on the combined data from the four oldest donors. The simulations were again performed using *rsimpop*, and the ABC regression steps were again performed using the R package *abc*[130]. In the case of non-neutral models, it is no longer the case that all the information provided by the genomic data, about the parameters of the model, is contained in sequence of inter-coalescent intervals (calculated from the phylogeny). Therefore, additional summary statistics (including clade size statistics) were used in the ABC steps.

A separate donor-specific posterior predictive p-value (donor-specific PPC) was computed for each donor (not only for the four oldest donors), based on the (approximate) multiple-donor posterior distribution on the combined data from the 4 oldest donors. In this case, the sample from each donor-specific posterior predictive distribution was generated by repeated sampling of parameter values from the multiple-donor posterior distribution, and then re-simulating the model (using *rsimpop*) conditional on the donor-specific sample size (number of single cell genomes) and donor age. As before, the posterior predictive p-value is computed from the upper tail-area probability under the distribution of the difference between the simulated chi-squared discrepancy and the observed chi-squared discrepancy[132].

The purpose of this donor-specific PPC is to decide if the observed phylogeny obtained from each donor is compatible with the simple non-neutral model (while taking account of our uncertainty about the parameter values in the model). If the p-value is close to zero, then the observed data is extreme (an outlier) compared to the data predicted under the simple non-neutral model. This is interpreted as evidence that the observed phylogeny is not compatible with the simple non-neutral model, and that more elaborate models need to be considered.

2.27 Phylofit estimation of selection coefficients

Written by Nick Williams

The algorithm *phylofit* was used to estimate the selection coefficients of known and unknown drivers in our phylogenies. *Phylofit* uses an efficient MCMC approach to model selection within a clade using the probability density of coalescence times and the population size trajectory. As such it can be thought of as a parametric adaptation of the *phylodyn* model.

The starting point for *phylofit* is Equation 1 in Lan *et al* 'An Efficient Bayesian Inference Framework for Coalescent-Based Nonparametric Phylodynamics'[144]:

$$P(t_1,..,t_n|N(t)) = \prod_{k=2}^{n} \binom{k}{2} \frac{1}{N(t_{k-1})} e^{-\int_{t_k}^{t_{k-1}} \binom{k}{2} \frac{1}{N(t_{k-1})} dt}$$

Where $\{t_k | k \in 1..n\}$ are the timings of the time ordered coalescences belonging to the driver mutation containing clade, $t_1$is the first coalescence of the expansion and t_n is the sampling time. These times are expressed as the interval between the event and the sampling time (assumed to be isochronous).

Substituting this formula for the cell count of the driver mutation containing clade $N(t)$ (in our case aberrant cell count refers to expanded clades both with and without known drivers) and performing the integral, eliminating terms that do not depend on overall population size, N , the trajectory midpoint, $t^{(m)}$, and the selective coefficient, $\hat{s} = \alpha s$, the following log-likelihood is derived:

$$L\left(t_1,..,t_n|\hat{s},t^{(m)},N\right) =$$

$$(n-1)\log(\mathrm{N}) + \sum_{k=2}^{n}\left(\log\left(1+\exp\left(\hat{s}\left(t_{k-1}-T+t^{(m)}\right)\right)\right)\right) -$$

$$\frac{1}{\hat{s}N}\sum_{k=2}^{n}\left(\binom{k}{2}\exp\left(\hat{s}\left(t_{k-1}-T+t^{(m)}\right)\right)\left(\exp\left(\hat{s}(t_{k-1}-t_k)\right)-1\right)\right) +$$

$$\frac{1}{N}\sum_{k=2}^{n}\left(\binom{k}{2}\hat{s}(t_{k-1}-t_k)\right)$$

Where recall the annualised selective coefficient is $S = \exp(\alpha s) - 1 = \exp(\hat{s}) - 1$
This central likelihood equation is into a Bayesian model with uniform priors on log(N), $\hat{s}$ and $t^{(m)}$.

$$\hat{s} \sim U(0.001, 2)$$
$$t^{(m)} \sim U(a, b)$$
$$\log 10(\mathrm{N}) \sim U(4, 6)$$
$$\boldsymbol{t} \sim Phylo(\hat{s}, t^{(m)}, N)$$

Here $Phylo$ is the probability distribution described by the log-likelihood function specified above.

Additionally, assuming unbiased sampling, we can optionally incorporate the number of sampled driver mutation containing colonies n_{mut} out of n_{tot} total colonies can be incorporated as an additional layer in the model:

$$n_{mut} \sim \mathrm{Binomial}\left(n_{tot}, \frac{1}{1 + \exp\left(-\hat{s}(T - t^{(m)})\right)}\right)$$

The parameters, a and b, setting the realistic range for the midpoint depend on whether the last component of the model is active and are detailed in the code.

The above models were coded in R and Rstan and inferred using the Rstan implementation of Stan's No-U-Turn sampler variant of Hamiltonian Monte Carlo method*. Models were fitted across three chains each with 20,000 iterations including 10,000 burn-in iterations.

The input data for this approach is an ultrametric tree. We obtain the ultrametric tree for this analysis using the methods already outlined. The code used to run *phylofit* can be found at (https://github.com/emily-mitchell/normal_haematopoiesis/7_phylofit/scripts/phylofit.R)

The *phylofit* algorithm was validated by assessing the correctness of the selection coefficient inference when the algorithm was run on single driver mutation clones with a known selective coefficient. The procedure was as follows:

Simulate population with initial division rate of 0.1 per day ($\alpha = 0.1$) until population has grown to the target equilibrium population size.

- Set symmetric division rate to 1 per year ($\alpha = 0.5/365$) and simulate neutral evolution until time T=5 years.
- Save the state of the simulation (*)
- Introduce the driver with the specified selection coefficient.
- If the driver lineage dies out before the sampling age is reached, or has less than 2% clonal fraction at the sampling age, then return to the saved state (*) and continue.

An unbiased sub-sample of cells is taken from the extant population of cells. The *phylofit* algorithm was then applied to the mutant clade in the sub-sampled simulated ultrametric phylogenetic tree.

The algorithm was found to recover the selection coefficients over a range of values of selection coefficient (**Fig. 2.13**).

*Stan Development Team (2020). "RStan: the R interface to Stan." R package version 2.21.2, http://mc-stan.org/.

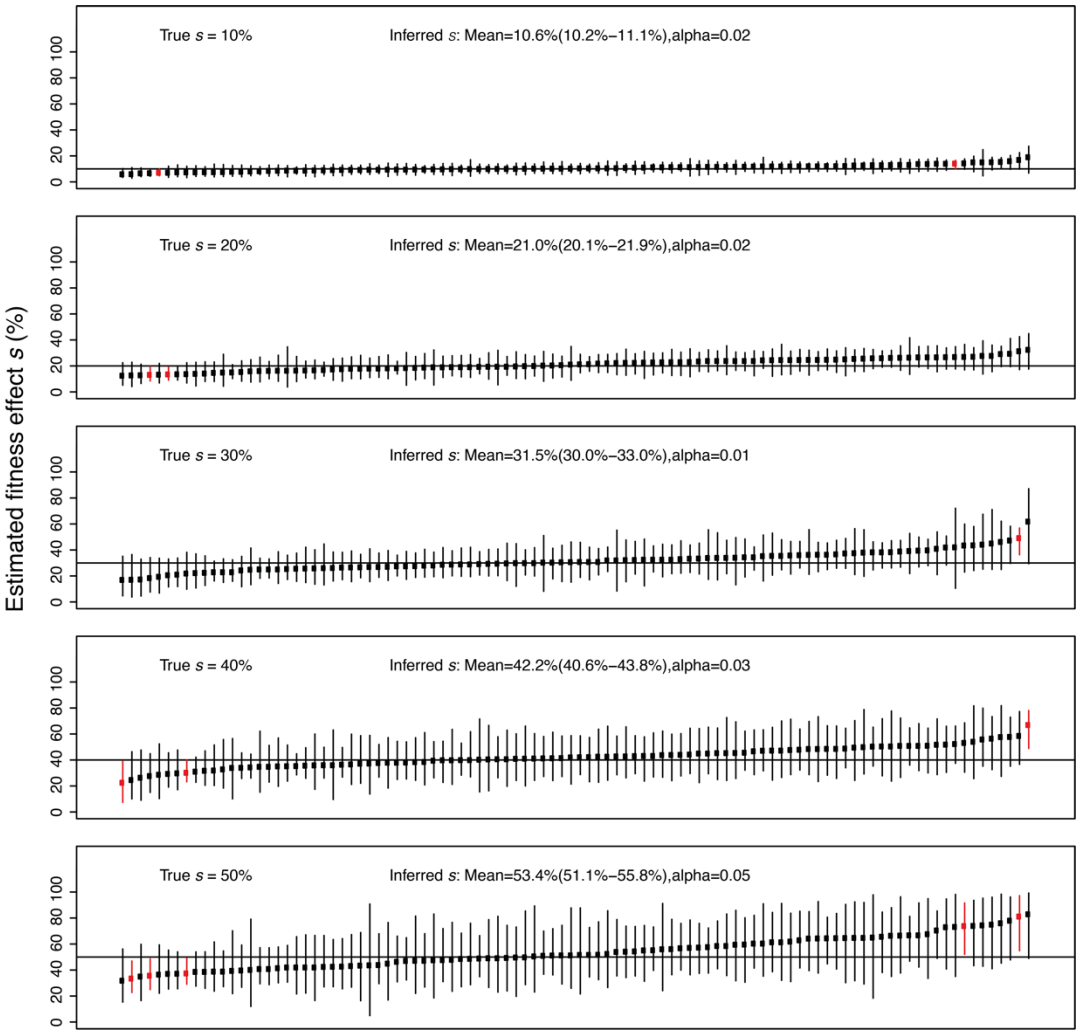


Fig. 2.13| Phylofit benchmarking. The inference of annualised fitness effect, *s*. The *phylofit* results (prior *s* range is 0-100% and log10(N) is 4 to 6) are shown for one hundred simulations for each of five values of *s* (=10%, 20%, 30%, 40% and 50%) and N=100,000 cells. The vertical lines show the 95% credibility intervals of the inferred selection coefficients with red lines highlighting instances where the true selection coefficient lies outside the 95% credibility interval ("alpha" is the proportion of such cases). The sample mean estimate of *s* and the corresponding 95% confidence interval are also shown. The benchmarking shows that on average the selection coefficient is accurately recovered with little bias. *Plots shown generated by Nick Williams.*

2.28 Analysis of Acute Myeloid Leukaemia (AML) genomes

Mutations in *ZNF318* and *HIST2H3D* were identified in recently described tumour WGS data from 263 patients with AML and MDS seen at Washington University School of Medicine in St. Louis[145]. Sequencing, initial processing using the *hg38* human reference genome, and full variant calling details for this dataset were described previously[145]. Briefly, identification of SNVs and indels in *ZNF318* and *HIST2H3D* was performed with *Varscan2* run in SNV and indel mode using custom parameters to enhance sensitivity (--min-reads2=3, --min-coverage=6, --min-var-freq=0.02, and --p-value=0.01), along with indel callers *Pindel* and *Manta* using default parameters. Variant calls identified via these approaches were merged and harmonized using a custom python script and annotated with VEP using *Ensembl* version 90. Only one potentially pathogenic variant was found in *ZNF318*. All identified variants are listed in **Table S4**, the majority are likely to be rare germline variants as the sequencing strategy did not incorporate a matched normal sample.

In addition there were no variants in *ZNF318* or *HIST2H3D* reported in 200 cases of AML in the TCGA dataset[146], compared to 51 cases with DNMT3A mutations. Similarly of 71 AML cases in the JAMA study[147] 23 cases had DNMT3A mutations but none had *ZNF318* or *HIST2H3D* mutations. Both studies used a tumour/normal WGS approach and reported all somatic variants identified as part of their supplementary datasets.

2.29 HDP signature extraction (chemotherapy dataset)

Mutational signature extraction was performed *de novo* and without priors using a Hierarchical Dirichlet Process (HDP, https://github.com/nicolaroberts/hdp). Samples were grouped by patient but no other clustering information was provided to HDP.

2.30 Number of cell divisions between HSCs and mature peripheral blood cells

I performed these calculations primarily for another manuscript[148], but for interest I have also outlined them here.

To explore the implications of the similar mutation burden between HSPCs and terminally-differentiated granulocytes the upper and lower bounds of the likely number of cell divisions between HSCs and mature peripheral blood cells in humans (q) were estimated. MacKey

(2001)[149] calculated *q* in mice to be between 17 and 19.5 using parameters for the HSC population size (*N*), the haematopoietic production rate (*HPR*) and the HSC differentiation rate (δ) and the following equation. If the average effective amplification (*A*) in cells numbers between HSCs and their mature progeny is given by $A = 2^q$ then:

$HPR = \delta N \times 2^q$

The required parameters have more recently been estimated in humans (**Table 2.4**). Human HSC population size estimates range from 25,000 to 1.3 million and the human HSC differentiation rate is estimated to be between 1 per 28 days and 1 per 4 years. Using these inputs, we have calculated that the number of cell divisions between HSCs and mature cells (*q*) lies between 24 and 35 in humans.

A division rate of 1 per 28 days can be calculated to be the upper limit for human HSC divisions given published estimates of 1.2 mutations acquired per cell division and that HSCs have previously been shown to accumulate approximately 16 mutations per year[15]. This upper limit for division rate represents the situation where mutations are only acquired as a direct result of cell division and not with time. However, the more credible estimates for *q* are likely to be towards the upper end of the calculated range, as when an HSC division rate of 1 per 28 days is used, we can calculate that HSC lineages would on average have undergone 383 divisions by the age of 35 (**Table 2.5**). This number of cell divisions by middle age is very different to the upper estimate of 18 in mice. It also seems implausible given the well described Hayflick limit of 50 divisions prior to cellular senescence in cultured cells *in vitro*[59], although it is not known if this limit applies to stem cells *in vivo*. However, granulocyte telomeres have been shown to decline with age, which suggests that telomerase is not active in HSCs, and therefore that the haematopoietic system is not immortal with regard to numbers of replications over the lifespan of an individual[58,150]. Given these lines of evidence, an HSC division rate closer to 1 per every 4 years would seem more likely.

A *q* of 35 would predict an additional 42 mutations in granulocytes compared to HCSs (35 x 1.2), if the time taken for differentiation to occur is negligible. This is close to the observed difference between the number of mutations in cord blood HSCs (mean 66 mutations) and

cord blood granulocytes (mean 109 mutations) in Abascal *et al*[148] and would also fit with the results they observed in adults.

Table 2.4: HSC parameter estimates

Parameter	Mouse	Human
HSC population size (N)	(1.12E+04, 2.24E+04)***1***	(4.4E+04, 2.15E+05)***2*** (2.5E+04, 1.3E+06)***3***
Haematopoietic production rate (cells/day)	(1.88E+08)***4***	(4.69E+11)***4***
HSC differentiation rate (per day)	0.069 (0, 0.2)***5*** 0.228 (0,0.59)***6*** 0.009 (LT-HSC), 0.45 (ST-HSC)***7***	
HSC cell division rate (per day)	(6.0E-03, 3.0E-02)***8*** (2.0E-02, 5.0E-02)***9***	Self-renewal divisions (1.7E-02, 1.7E-03)***2*** (6.8E-04, 3.0E-02)***3***
Number mutations acquired per cell division		(1.2)***2***

1 Abkowitz *et al* 2002[151]
2 Lee-Six *et a*l 2018[15]
3 Watson *et al* 2020[152]
4 MacKey *et al* 2001[149]
5 Bradford *et al* 1997[153]
6 Cheshier *et al* 1999[154]
7 Busch *et al* 2015[105]
8 Kaschutnig *et al* 2015[155]
9 Foudi *et al* 2009[156]

Table 2.5: Calculation of the number of stem cell divisions between HSCs and mature cells

	Mouse	Human
Weight used for estimates (kg)	0.025	70
Lifespan used for estimates (years)	2	70
Age at middle age (years)	1	35
Haematopoietic production rate (HPR) - cells/day	1.88E+08	4.69E+11
HSC population size (N)	(1.12E+04, 2.24E+04)	(2.50E+04, 1.30E+06)
HSC differentiation rate (per day)	(6.00E-03, 2.28E-01)	(6.80E-04, 3.00E-02)
Effective amplification (A)	(3.67E+04, 2.79E+06)	(1.20E+07, 2.76E+10)
Number of cell divisions between HSCs and mature cells (q)	(1.52E+01, 2.79E+06)	(2.35E+01, 3.47E+01)
HSC cell division rate (per day)	(6.00E-03, 5.00E-02)	(6.80E-04, 3.00E-02)
Number of cell divisions by middle age	(2.19E+00, 1.83E+01)	(8.69E+00, 3.83E+02)

2.31 Data availability

Ageing dataset

All scripts and some smaller data matrices are available on github (https://github.com/emily-mitchell/normal _haematopoiesis). Raw sequencing data is available on the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/home; accession number EGAD00001007851). The main data needed to reanalyse / reproduce the results presented is available on Mendeley Data (https://data.mendeley.com/datasets/np54zjkvxr/1).

See below for a guide to what is available on Mendeley Data.

dNdS_input folder

Contains all raw input files for the dN/dS analysis.

Filtering_output_XXXX folders (one for each individual)

Contains four files:

a) annotated_mut_set_XXXX_01_standard_rho01

This is an R data object and is uploaded into an R workspace using load()

The genotype matrix used for MPBoot tree building is available in the matrix: filtered_muts$Genotype_shared_bin

The dna strings used as input for MPboot are available in the vector: filtered_muts$dna_strings

The annotated variant calls with tree node information are available in the matrix: filtered_muts$COMB_mats.tree.build$mat

The genotype matrix of mutations calls per sample is available in: filtered_muts$COMB_mats.tree.build$Genotype_bin

Information on whether the variant is an SNV or indel is available in: filtered_muts$COMB_mats.tree.build$mat$Mut_type

A summary of total numbers of shared and private SNVs and indels is available in: filtered_muts$summary

b) XXXX_sensitivity

This file contains information on the sensitivity of SNV and Indel calls per sample.

c) tree_XXXX_01_standard_rho01.tree

The raw tree with branch lengths equal to number of mutations assigned (without adjustment for sequencing coverage).

metadata_matrix folder

Contains file "Summary_cut.csv" which records metadata on each sample in the dataset including cell_type sorted, sequencing depth, sequencing_platform, SNV burdens, indel burdens and telomere length.

Chemotherapy dataset

Data is not yet publicly available as the analysis is still being finalised.

2.32 Code availability

Ageing dataset

Code is available on github:
https://github.com/emily-mitchell/normal_haematopoiesis/

Chemotherapy dataset

Complete code is not yet publicly available as the analysis is still being finalised.

Chapter 3: Age-related change in haematopoietic stem cells

3.1 Introduction

Many changes in HSCs and their niche have been hypothesised to contribute to age-related loss of function in the haematopoietic system. Age-associated phenotypes include anaemia, loss of regenerative capacity, especially in the face of insults such as infection, chemotherapy or blood loss; and an increased risk of blood cancer. Multiple lines of evidence from mouse models, suggest that cell intrinsic factors may be relatively more important in HSC ageing than in the age-related changes of other organ systems. Firstly, some HSCs in old mice do retain a youthful function, showing there is distinct heterogeneity of ageing with the population[65,66,157]. Secondly, transplantation of aged murine HSCs into young mice does not restore youthful function[65,69,157]. Thirdly, systemic rejuvenation interventions can restore youthful function in some murine organ systems but not aged HSCs[158]. Very little work to date has looked at age-related intrinsic changes in human HSCs at the single cell level. The experimental approach developed for this project allowed a number of cell intrinsic changes with age to be investigated in a large sample of single HSCs across the age range of the full human lifespan.

Key questions to be addressed in this chapter

1. Does human HSC mutation rate accelerate in old age?
2. What is the trajectory of telomere loss with age in human HSCs?
3. What is the impact of ageing on mature cell output of single human HSCs?
4. How does the human HSC population size and structure change with age?

These questions were investigated by whole genome sequencing and immunophenotyping single HSC-derived colonies from a range of sources and donor ages. I performed all the laboratory work for this 'ageing' study. I also performed all the analysis presented in this chapter, other than the structural variant analysis, which was carried out by Hyunchul Jung and LOY variant calling which was carried out by Tom Mitchell. I benefitted greatly from scripts and functions for custom variant filtering and tree building and plotting developed by

Mike Spencer Chapman[3] and Nick Williams[4] which I adapted for this work. The HSC population modelling work I performed was completely reliant on the R package *rsimpop*, an HSC population simulator developed by Nick Williams, and a method of posterior predictive checking developed by Kevin Dawson. The results below (along with those in **Chapter 4**) have been accepted for publication in Nature. The phenotyping analysis described in **Section 3.6** was not included in that publication and will be incorporated into a subsequent body of work.

3.2 Clinical information and samples

Samples were obtained from ten individuals aged between 0 and 81 years (**Table 3.1**), and utilised as shown in the experimental approach in **Fig. 3.1a**. All subjects were haematologically normal. Of note, one subject had inflammatory bowel disease (Crohn's disease) treated with azathioprine (KX002, 38-year male) and one had selenoprotein deficiency[159], a genetic disorder not known to impact HSC dynamics (SX001, 48-year male). None of the individuals had been exposed to chemotherapy. The source of stem cells was cord blood for the two neonates, and bone marrow and/or peripheral blood for adult donors (**Fig. 3.1b**). Bone marrow samples were obtained peri-mortem, allowing sampling of large volumes (50-80ml) from multiple vertebrae.

Table 3.1: Clinical information normal ageing cohort

Donor ID	Age	Sex	Clinical information	Cause of death	Tissue source[1]	Total cells WGS
CB001	0	F	Normal	NA	CB	216
CB002	0	F	Normal	NA	CB	390
KX001	29	M	Normal	Trauma	BM / PB	408
KX002	38	M	Crohn's disease	Intracranial haemorrhage	BM	380
SX001	48	M	Selenoprotein deficiency	NA	PB	363
AX001	63	M	Normal	NA	PB	361
KX007	75	M	Normal	Intracranial haemorrhage	BM	315
KX008	76	F	Normal	Intracranial haemorrhage	BM	367
KX004	77	F	Normal	Trauma	BM	451
KX003	81	M	Normal	Trauma	BM / PB	328

[1]CB = cord blood, BM = bone marrow, PB = peripheral blood

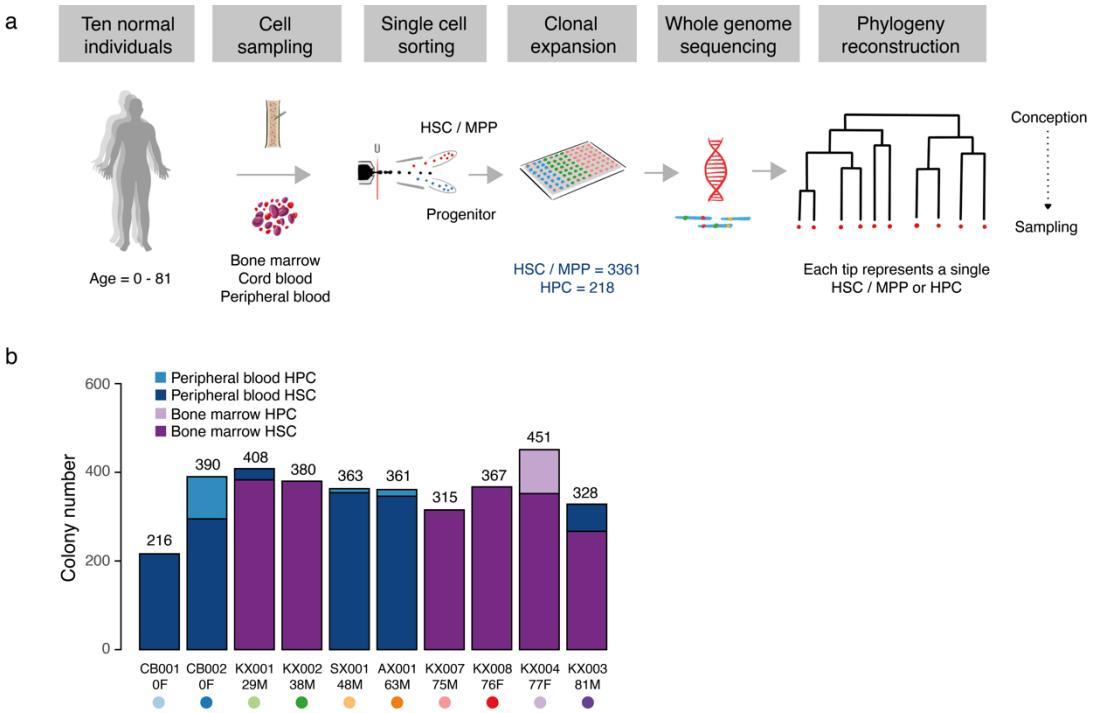


Fig. 3.1| Experimental approach and summary of samples. a, Experimental approach. **b,** Barplot showing the age, sex and numbers of colonies sequenced from each tissue and cell type for each donor in the study.

3.3 Overview of experimental approach

Laboratory work

For all individuals, single immunophenotypic haematopoietic stem cell/multipotent progenitors (HSC/MPPs: Lin-, CD34+, CD38-, CD45RA-) were flow-sorted[24] into 96 well plates and grown in liquid culture (**Fig. 2.1, Chapter 2**). Overall, 42-89% of sorted HSC/MPPs produced colonies, meaning that the sequenced colonies were a representative sample of the HSC/MPP population in each individual (**Fig. 2.2a**). Analysis of cell surface markers showed no difference between the immunophenotypic characteristics of 'sequenced' and 'not-sequenced' cells, further supporting the view that the colony growth step did not result in measurable bias in terms of cell type subsets sequenced (**Fig. 2.2b**). For four individuals, haematopoietic progenitor cells (HPCs: Lin-, CD34+, CD38-) were also sorted to allow comparison of mutation burden between HSC/MPPs and HPCs.

After three weeks in liquid culture, immunophenotyping was performed on the mature cells of colonies greater than approximately 3000 cells in size (**Fig. 2.3a**). This provided a measure

of proliferative potential (colony size) and lineage output (mature cell types produced). After another week in culture, 16 mature cells from a subset of 96 colonies were single cell sorted for scRNA seq. The scRNAseq data is currently being analysed by Lori Kregar and is not presented in this thesis.

Whole genome sequencing

Whole-genome sequencing was performed on DNA extracted from residual cells post immunophenotyping or on DNA extracted from the whole colony for those between 200 and 3000 cells in size. Average sequencing depth was 14X and was performed on 224-453 colonies per individual. In total 17 colonies were excluded due to low coverage, 34 as technical duplicates and 7 that were derived from more than a single cell (**Fig. 2.5**). The final dataset contained whole genomes from 3579 colonies, of which 3361 were single HSC/MPP-derived and 218 were single HPC-derived.

DNA from a subset of colonies from these individuals is also undergoing NEB enzymatic methylation sequencing. The methylation data is currently being analysed by Joe Lee and Lori Kregar and is not presented in this thesis.

Data analysis and validation

Variant calling and filtering strategies followed largely established approaches and are described in detail in **Chapter 2**. To allow a valid analysis of mutation burden across our cohort, raw SNV and indel mutation burdens were corrected for sequencing depth using asymptotic regression (**Fig. 2.7**). The quality of variant filtering was confirmed by comparing single base substitution spectra of our individuals with previously published data[3,53,86], which were highly concordant (**Fig. 2.6b**). Spectra of shared mutations (variants called with high confidence) were also compared to the spectra of private mutations (only found in a single sample and therefore called with lower confidence). The spectra of shared and private mutations were also reassuringly concordant, suggesting minimal artefactual calls in the dataset (**Fig. 2.6a**). Phylogenetic trees for each individual were constructed from the patterns of shared and unique somatic mutations in their sampled cells (**Chapter 2, Fig. 2.8)**.

3.4 HSC mutation accumulation over life

Previous work supported the view that single nucleotide variant accumulation in HSCs is linear over life[38,53,87]. However, Welch *et al* performed exome-sequencing meaning resolution was very low (fewer than 15 variants called per HSPC) and the oldest individuals in both the Osorio *et al* and de Kanter *et al* studies were less than 65. Therefore, the question of whether mutation accumulation may accelerate in HSCs in old age remained largely unanswered at the start of this project. Accurate estimates of indel burden, structural variant burden and copy number variant burden in representative samples of normal HSCs across the lifespan had also not previously been made.

Single nucleotide variant burden

Single nucleotide variants accumulated in the HSC/MPPs dataset linearly over life from birth into old age. Linear mixed effects modelling was used to estimate the SNV mutation rate in HSC/MPPs at 16.8 substitutions/cell/year ($CI_{95\%}$=16.5-17.1; **Fig. 3.2a**). There was no elevation in mutation rate in old age as assessed by breaking down the linear regression into three age ranges (0-29, 29-63, 63-81), with almost identical rates estimated for all three age brackets (**Chapter 2**).

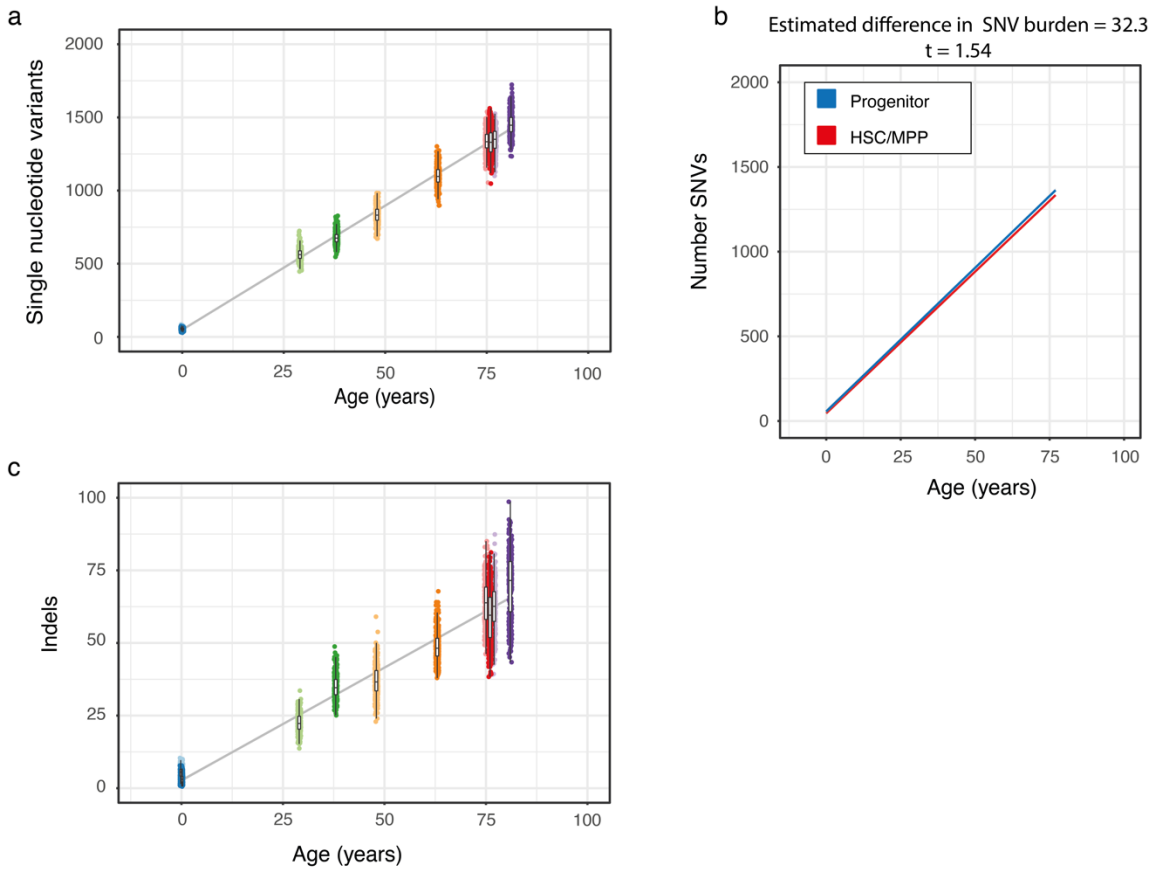


Fig. 3.2| SNV and indel burden in normal HSC/MPPs. a, Burden of single nucleotide variants (SNVs) across the donor cohort. The points represent individual HSC/MPP colonies (n = 3361) and are coloured by donor. The boxes overlaid indicate the median and interquartile range and the whiskers denote the range. The grey line represents a regression of age on mutation burden, with 95% CI shaded. **b,** Regression of number of single nucleotide variants in HSCs (red line) compared to HPCs (blue line). Grey shading indicates the 95% CI. The estimated difference in burden, together with the t-value is above the plot. The t-value of 1.54 demonstrates non-significance of the difference. **c,** Burden of small indels across the donor cohort. The points, boxes and line are depicted as in **a**.

The data presented here are also consistent with previous studies that have shown an elevated mutation rate during embryonic development and foetal life (prenatally)[3,53]. The mean mutation burden across HSC/MPPs from the two cord blood samples was 55 SNVs/cell. This equates to a rate of 75 SNVs/cell/year in utero, much higher than the rate observed in postnatal life. Previously published data on whole genome sequencing single HSPC-derived colonies from two human foetuses allow us to refine the pre-natal mutation rates further[3]. A mean mutation burden in HSPCs at 8 weeks' gestation of 25 SNVs/cell, and at 18 weeks of 42 SNVs/cell was observed. These data show that the rate of mutation acquisition slows considerably during development, from an average rate of 3.2 mutations per week in the first

8 weeks to an average rate of 1.6 mutations per week between weeks 8 and 18. The mutation rate likely also slows considerably in the latter half of gestation when compared to our cord blood mutation burden of 55. This provides an estimate of an average rate of 0.65 mutations per week between weeks 18 and 38, which is approximately double the rate calculated for post-natal life of 0.32 mutations per week.

The study on foetuses is also helpful in demonstrating there is a rate of mutation acquisition per cell division of < 0.9 after the first 3 cell divisions in the zygote. Although this rate of mutation accumulation per cell division is low, due to the very much higher rates of cell division in early development, it may account for the overall higher rates of mutation acquisition observed.

Mixed effects modelling found no significant difference in mutation burden between HSC/MPPs and HPCs (**Fig. 3.2b**). The estimate of the increased number of SNVs in HPCs versus HSC/MPPs was approximately 30 mutations (with a non-significant t-value of 1.54). This is similar to the number of excess mutations observed in granulocytes compared to HSC/MPPs in a recent study describing a novel single molecule sequencing approach 'Nanoseq' (which found no significant difference in mutation burden between HSCs and granulocytes). The Nanoseq study used a subset of the HSC/MPP mutation burden data presented here in a comparison with bulk granulocyte mutation data generated using Nanoseq. As part of work for that study I estimated that at least 30 cell divisions are required to generate granulocytes from HSCs based on current parameter estimates (**Chapter 2**). These results therefore support the view that very few mutations accumulate per cell division as previously shown in analyses of embryonic mutation rate discussed above[3,86] (most likely less than 1), with the majority of mutations accruing in a time dependent manner.

Indel burden

Indel burden was over 10-fold lower than the SNV burden in HSCs and showed a greater variability in cells from the same individual. In an identical approach to that used for SNVs we estimated that indels accumulate at a rate of 0.71 indels/cell/year postnatally, again showing no obvious acceleration of rate in old age (**Fig. 3.2c**). As with SNVs, there was a higher rate of

indel accumulation during *in utero* development, with a mean burden of 2.68 indels per HSC at birth.

Structural variant burden

Structural variants were rare, with only 1-17 events observed in each individual (**Fig. 3.3a**).

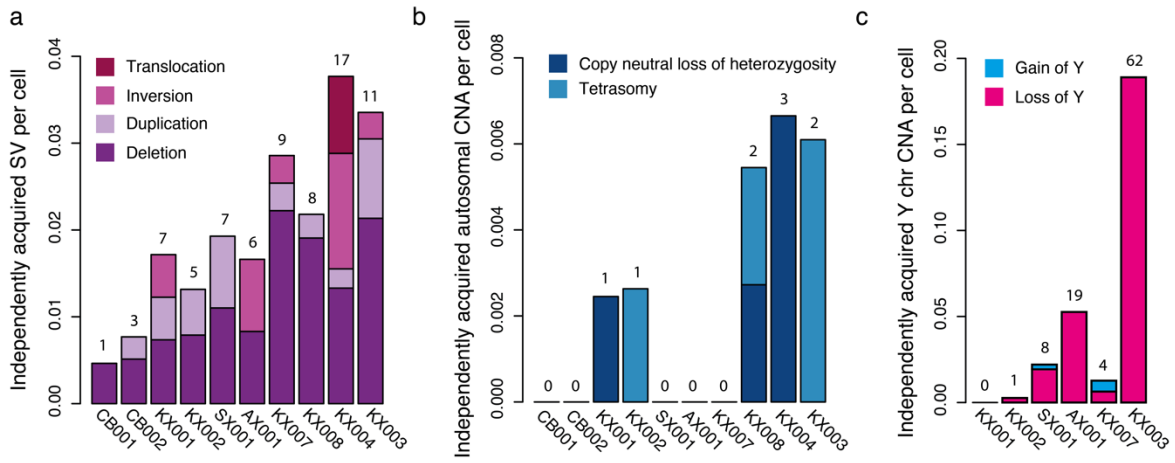


Fig. 3.3| SNV and CNV burden in normal HSC/MPPs. a, Barplot of the number of independently acquired structural variants (SVs) per colony sequenced in each donor. The absolute number of SVs is at the top of each bar. **b,** Barplot of the number of independently acquired autosomal copy number aberrations (CNAs) per colony sequenced in each donor. The absolute number of CNAs is at the top of each bar. **c,** Barplot of the number of independently acquired Y chromosome copy number aberrations sequenced in each male donor. The absolute number of CNAs is at the top of each bar.

Most events were deletions and burdens correlated with age (**Table S5**). Structural variants could be layered on the phylogenies and were found to have occurred through life. One was identified as having occurred during the first cell division, another was timed to in utero, while others could be timed to the latter half of life (**Fig. 3.4**).

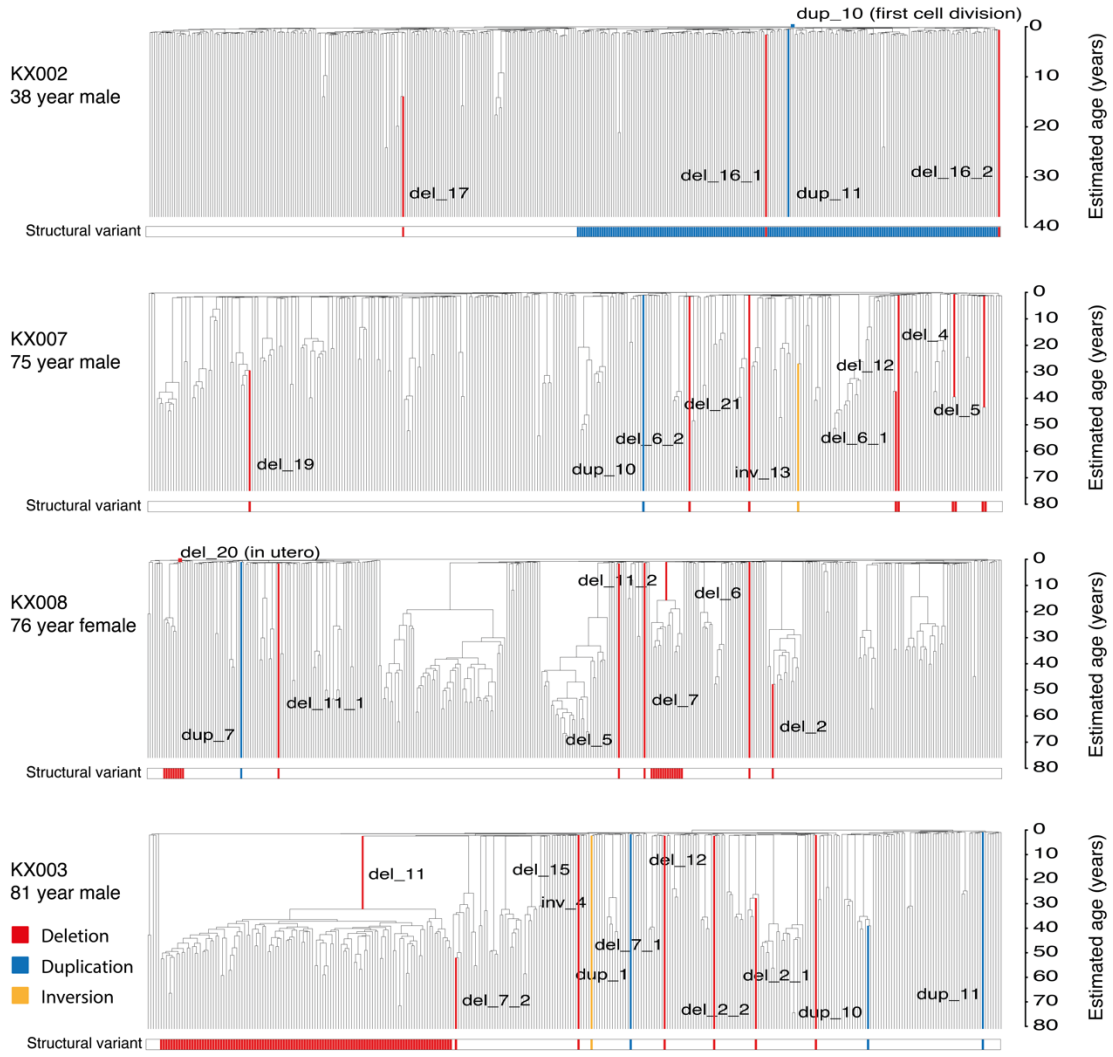


Fig. 3.4| Structural variants and CNAs layered on phylogenies. Phylogenies depicted for the individuals with clonally expanded structural variants (SVs). The bar at the bottom highlights cells with one of the three classes of structural variant. The exact variant breakpoints can be found in **Table S5**.

Copy number variant burden

Autosomal copy number changes were rare at all ages and comprised either copy-neutral loss of heterozygosity events or tetrasomies (**Fig. 3.3b**). No X chromosome copy number changes were observed. In contrast, Y chromosome copy number changes were frequent in males, increasing with age as previously shown[160] (**Fig. 3.3c**). Loss of Y (LOY) events were much more frequent than gain of Y events and will be discussed in more detail in **Chapter 4**.

Summary mutation burden

1. SNVs accumulate in HSCs linearly over post-natal life at a rate of 16.8 SNVs/cell/year.

2. Indels accumulated in HSCs linearly over post-natal life at a rate of 0.71 indels/cell/year.
3. A higher SNV and indel mutation rate was observed during *in utero* development.
4. There was no acceleration of SNV or indel mutation rate in old age.
5. Structural variants were rare in normal HSCs, but burdens correlated with age.
6. Autosomal copy number changes were rare at all ages.
7. No X chromosome changes were identified.
8. LOY events were frequent in males and increased with age.

3.5 HSC telomere length changes over life

Telomere attrition over life

Telomere lengths could be assessed using the algorithm *telomerecat* for 1505 HSC/MPP colonies from seven individuals that were sequenced on Hiseq X10[124]. As previously reported[42,125,161], telomere lengths decreased steadily with age, at an average attrition rate of 30.8 bp per year in adult life ($CI_{95\%}$=13.2-48.4) (**Fig. 3.5a**), which is close to published estimates of 39bp/year from bulk granulocytes[42]. The bulk granulocyte study observed higher rates of telomere loss in the first 6 months post birth which would fit with the relatively longer telomeres observed in the cord blood HSC/MPP data. When comparing the adult HSC telomere lengths to those in cord blood, there is an excess telomere loss of approximately 1800 bp in early life (from published bulk data[42] this is likely to be within 6 months post birth). A similar rapid loss of telomere length in early life has been found in longitudinal studies of baboon granulocytes, which lost 2000-3000bp in the first year post birth)[162]. This more rapid telomere attrition in very early postnatal life likely reflects a period of increased stem cell turnover due to continued expansion of the HSC population in this period. This explanation would fit with the observed HSC population trajectories generated from the phylogeny data in **Section 3.7**, which show the HSC population expands rapidly *in utero* before stabilising at adult levels around the time of birth.

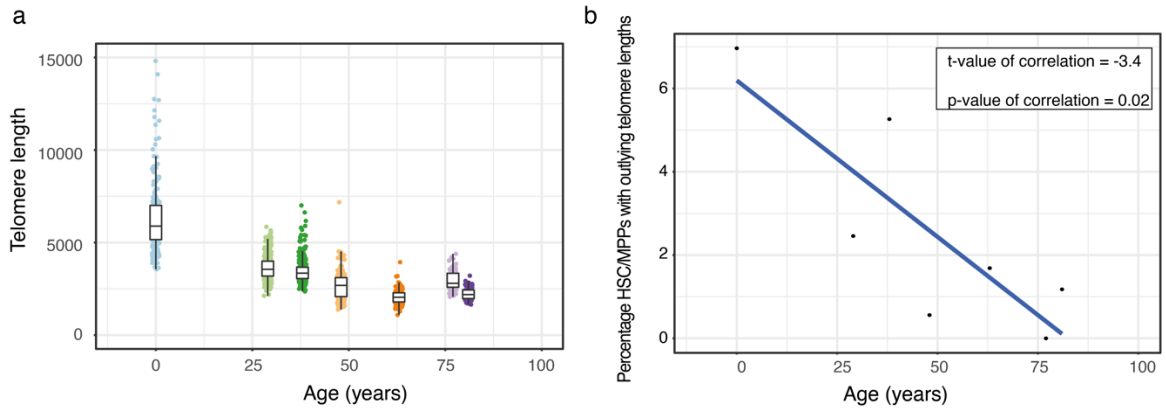


Fig. 3.5| Telomere attrition over life. a, Telomere length across the donor cohort, including only those samples sequenced on the HiSeq X10 platform. Each point represents a single HSC/MPP colony. The boxes overlaid indicate the median and interquartile range and the whiskers denote the range. Two outlying points for CB001 are not shown (telomere lengths 16,037bp and 21,155bp). **b,** Plot showing the percentage of HSC/MPP cells that have outlying telomere lengths per individual. Outliers were identified using the Interquartile Range criterion. There were no outliers with shorter than expected telomeres in any individuals, such that this data only reflects the percentage of cells with longer than expected telomeres. The blue line shows a regression of percentage outlying telomere lengths with age. This shows a significant negative correlation (t-value and p-value shown).

Telomere length distributions

The telomere data presented is highly novel, as sequencing single-cell derived colonies allows estimation of the variance and distribution of telomere lengths among cells in a population with a resolution that is not possible for bulk populations. In cord blood and adults aged < 65, a small proportion of HSC/MPPs had unexpectedly long telomeres. In contrast, there were no cells with unexpectedly short telomeres. The proportion of cells with outlying telomere lengths reduced in frequency with age (**Fig. 3.5b**). Given that telomeres shorten at cell division, these outlier cells have likely undergone fewer historic cell divisions. The finding of a population of less frequently dividing HSCs in humans would be in keeping with a previously described population of infrequently dividing HSCs in the mouse[74,156,163]. The data presented in **Fig. 3.5b** would suggest that the more dormant HSC population present in cord blood and younger adults is either lost or becomes relatively much rarer in elderly individuals. This loss of a more 'youthful' reserve population of HSCs in the elderly could in part explain the reduction in haematopoietic regenerative capacity with age. However, one caveat to the data presented is that it is based on a relatively small number of individuals, making it difficult to draw completely definitive conclusions.

Summary telomere lengths

1. Telomere lengths decreased with age at a rate of 30 bp per year in adult life.
2. Analysis of telomere length distributions identified a small proportion of HSC/MPPs with unexpectedly long telomeres, providing evidence for a population of relatively dormant HSCs in humans.

3.6 Mature cell output

Two hallmarks of ageing in the human haematopoietic system suggest that loss of proliferative potential occurs at some level of the haematopoietic hierarchy with age. Firstly, bone marrow cellularity universally declines from >90% in young childhood to < 20% by old age[164,165]. Interestingly, loss of bone marrow cellularity is not observed in laboratory mice kept in sterile conditions, but can be recapitulated if mice are repetitively exposed to agents that mimic infective stimuli[75]. Secondly, elderly humans commonly develop mild cytopenias of uncertain significance with age, and over the age of 65-70 are typically unable to tolerate intensive chemotherapy due to the risk of prolonged cytopenias and infection. In addition to decline in replicative function with age, studies of HSC ageing in mouse models have consistently demonstrated loss of lymphoid output with age, often described as a myeloid bias[71]. The extent to which this phenomenon is mirrored in humans has only been addressed in two studies to date[72,166].

In order to attempt to investigate the question of whether there is loss of HSC proliferative potential and or changes in differentiative output with age, I undertook immunophenotyping of single HSC and HPC cell derived colonies. Larger colonies from the individuals included in the whole genome sequencing study (typically colonies > approximately 3000 cells), underwent immunophenotyping analysis (**Fig.2.3**) to allow assessment of change in proliferative potential (colony size) and lineage output (mature cell types produced) with age. Immunophenotyping data included here was also obtained for two normal individuals not included in the whole genome sequencing (WGS) cohort (KX009 and KX010). In addition, immunophenotyping was performed on KX004 colonies from PB and spleen (SPL) that were not included in the WGS cohort. In total therefore, the immunophenotyping dataset is larger

than the WGS dataset, comprising 4478 single cell derived colonies across a range of ages and tissue types: CB (539), BM (3008), PB (642) and SPL (288). The majority of colonies were HSC derived (total 4127, **Table 3.2**) but some HPC derived colonies were also included (total 350, **Table 3.3**).

Table 3.2: Number HSC-derived colonies immunophenotyped by sample type

Donor ID	Age/Sex	CB	BM	PB	SPL
CB001	0 F	155	0	0	0
CB002	0 F	384	0	0	0
KX001	29 M	0	384	25	0
KX002	38 M	0	384	0	0
SX001	47 M	0	0	96	0
KX010	60 F	0	439	0	0
AX001	63 M	0	0	270	0
KX009	63 M	0	439	0	0
KX007	75 M	0	384	0	0
KX008	76 F	0	343	0	0
KX004	77 F	0	377	192	288
KX003	81 M	0	258	59	0

Table 3.3: Number HPC-derived colonies immunophenotyped by sample type

Donor ID	Age/Sex	CB	BM	PB	SPL
CB001	0 F	0	0	0	0
CB002	0 F	96	0	0	0
KX001	29 M	0	0	0	0
KX002	38 M	0	0	0	0
SX001	47 M	0	0	7	0
KX010	60 F	0	79	0	0
AX001	63 M	0	0	13	0
KX009	63 M	0	49	0	0
KX007	75 M	0	0	0	0
KX008	76 F	0	0	0	0
KX004	77 F	0	0	37	69
KX003	81 M	0	0	0	0

Colony size

Colony size is used here as a surrogate for proliferative potential. In keeping with expectations of relative proliferative potential, we observed that HSC-derived colonies were significantly larger than HPC- derived colonies (Welch's t-test p-value < 2.2E-16; **Fig. 3.6a**). This finding was confirmed both overall, and when comparisons were made within individuals for whom data was available for both cell types within a single sample type. HSC-BM colonies were significantly larger than HSC-CB, HSC-PB and HSC-SPL-derived colonies (Welch's t-test p-values = 0.00026, < 2.2E-16, < 2.2E-16 respectively; **Fig. 3.6b**). Again, the findings of this analysis are the same both overall, and within each individual.

In keeping with the observed size difference between HSC and HPC-derived colonies, there was also an observed difference in colony size by phenotype (**Fig. 3.6c**). Bi- or tripotent colonies (those containing two or more mature cell types; **Fig. 2.3b**) were significantly larger than their unipotent counterparts. For example, erythroid-only colonies were smaller than colonies containing both erythroid and myeloid cells. And granulocyte-only or monocyte-only colonies were smaller than colonies containing both monocytes and granulocytes. Of the

myeloid colonies, those containing predominantly monocytes were larger than those containing predominantly granulocytes. These findings suggest that there is heterogeneity of both lineage potential and replicative potential within the HSC/MPP compartment, with HSCs that show more lineage specificity also having reduced replicative potential.

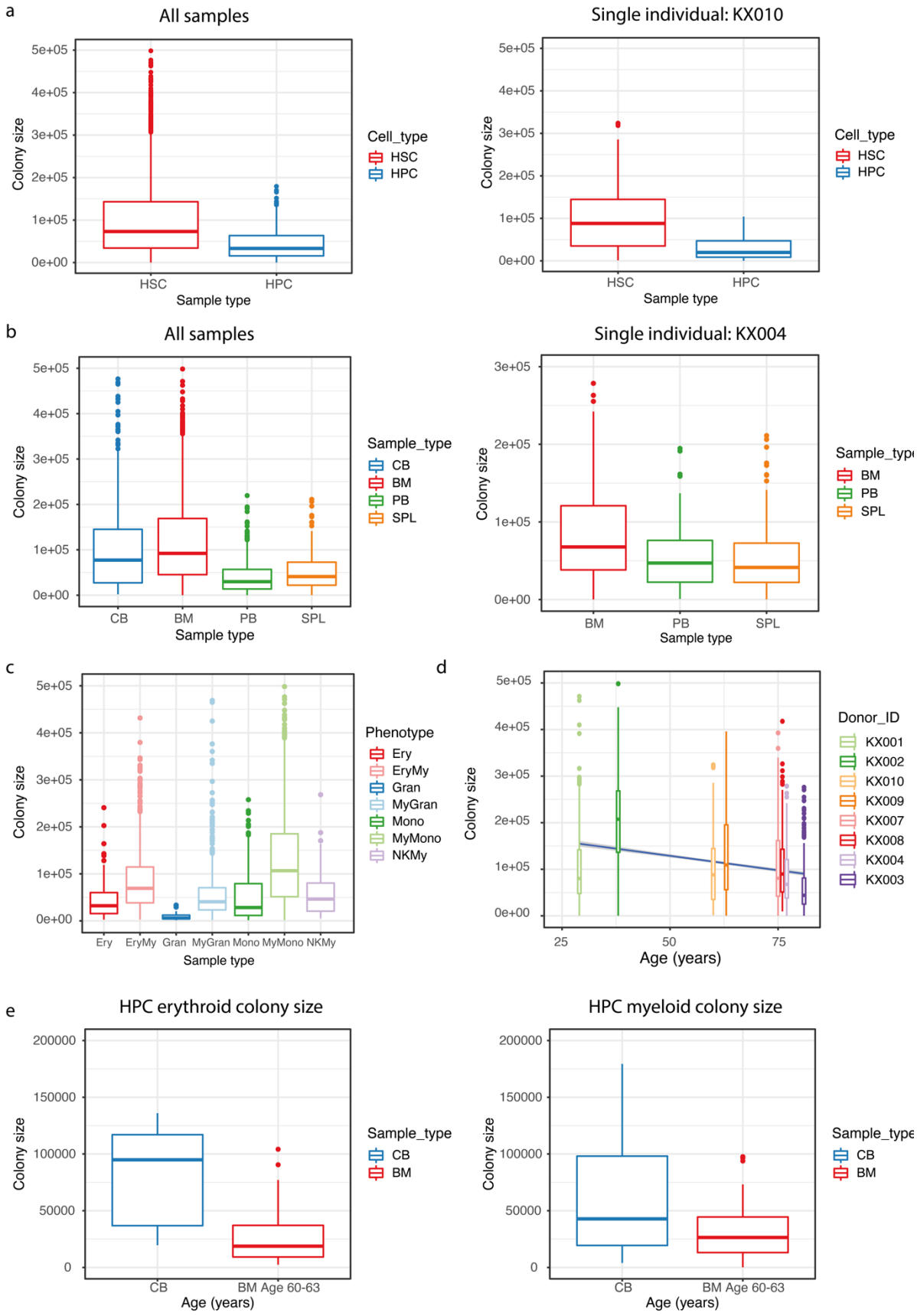


Fig. 3.6| Colony size. a, Boxplots comparing total size of HSC and HPC-derived colonies for all samples (left) and a single individual (KX010, right). **b**, Boxplots comparing total size of CB, BM, PB and SPL derived HSC colonies for all samples (left) and a single individual (KX004, right). **c**, Boxplots

comparing total size of HSC-derived colonies by colony phenotype. Ery = erythroid cells only; EryMy = erythroid cells and myeloid cells (any combination of monocytes and granulocytes); Gran = granulocytes only; MyGran = granulocytes and monocytes (granulocytes predominant); Mono = monocytes only; MyMono = monocytes and granulocytes (monocytes predominant); NKMy = NK cell and myeloid cells (any combination of monocytes and granulocytes). **d**, Boxplots showing total colony size by age for BM HSC-derived colonies. The blue line represents a regression of age on colony size, with 95% CI shaded. **e**, Boxplots showing total colony size for HPC erythroid colonies (left) and HPC myeloid colonies (right) from cord blood or older BM (aged 60-63).

Due to the fact that colony size differs by both cell type (HSC vs HPC) and sample type (CB, BM, PB and SPL), when looking for changes in colony size (or proliferative potential) with age the analysis has been restricted by both cell type and sample type. When analysis is restricted to BM-HSC derived colonies (for which the most complete data with age is available) there is a significant decrease in colony size with age (**Fig. 3.6d**). A linear model fitted to this data found a reduction in colony size of 1688 cells per year of life ($CI_{95\%}$=531-2847; t-value -2.9), predicting an BM HSC colony size of 222000 at birth. This significant reduction in colony size with age is largely driven by a single young individual, KX002. It is possible his underlying diagnosis of Crohn's disease, and associated chronic inflammation, may have played a role in driving a more proliferative HSC phenotype in this individual. However, even when the analysis is restricted to the oldest six individuals, a similar trend to reduced colony size with age is also observed, with a reduction in colony size of 1640 cells per year of life ($CI_{95\%}$=109-3171; t-value -2.1) which would support the findings across the whole age-span. Nevertheless, data from more young individuals, ideally including children, is required to more conclusively answer the question of whether there is evidence for loss of HSC proliferative potential with age.

Of note, manual observation of colonies from the individual with selenoprotein deficiency found that they were very small compared to normal individuals (such that it was only possible to immunophenotype a comparatively small fraction). This was likely due to the fact that they were grown under normoxic conditions *in vitro*, and therefore exposed to a much higher oxygen concentration than would be present in the bone marrow niche (known to be a hypoxic environment). Selenoprotein deficiency is an extremely rare condition (less than 10 cases diagnosed worldwide) leading to multisystem defects caused by the almost complete

absence of the 25 known human selenoproteins[159]. It results in azoospermia, muscular dystrophy, aortopathy, photosensitivity and increased cellular reactive oxygen species. The higher oxygen concentration experienced during *in vitro* growth almost certainly leads to increased intracellular reactive oxygen species, resulting in increased levels of apoptosis and the failure of cell proliferation we observed.

Although the current HPC immunophenotyping dataset is limited, it provides evidence that there is a more significant decline in HPC (rather than HSC) proliferative potential with age (**Figure 3.6e**). When colony size of both erythroid and myeloid HPC colonies from cord blood are compared to erythroid and myeloid HPC colonies from donors KX010 (aged 60) and KX009 (aged 63), a significant decrease in colony size with age is observed (Welch's t test p-value = 1.2E-09). Fewer HPCs underwent immunophenotyping from the older donors, largely because manual assessment of colony size using a microscope revealed the HPC-derived colonies (particularly myeloid HPC colonies) were too small. In comparison many of the myeloid HPC-derived colonies from cord blood were a similar size to HSC colonies. Our observation fits with other recent work, which found that erythroid colonies (that are more likely to be HPC than HSC-derived) decline in size with age[167]. It seems plausible that the probable decline in replicative potential observed at the HPC level in this assay, could explain the loss of bone marrow cellularity and function of the haematopoietic system as a whole with age. It will be exciting to more thoroughly characterise loss of HPC replicative potential with age in future work. Of particular importance will be to assess colony size of paediatric BM-derived cells with adult BM-derived cells, as the current dataset is limited to comparisons with CB which may be more analogous to adult PB. Despite this, the inferences on loss of progenitor size are likely valid as the results in adults would predict that HPC colonies from paediatric BM would be larger than those from CB.

Across the set of colonies for which we had available data (CB001, KX001, KX002, AX001, KX004, KX003), there was no association between colony size and telomere length. This is likely due to the fact that colony size is more heavily influenced by the differentiation state of the sorted cell (LT-HSC, ST-HSC, MPP), than telomere length.

Colony phenotype

Single cell derived colony phenotypes can be stratified by potency according to how many mature cell types are produced (**Fig. 2.3b**). Colony phenotype was found to differ by seeding cell type (**Fig. 3.7a**), with HSC-derived colonies predominantly containing a mixture of granulocytes and monocytes and HPC-derived colonies containing predominantly erythroid cells. A similar trend was observed by sample type, with BM- derived HSC colonies being predominantly myeloid, and CB, PB and SPL derived colonies being more likely to contain erythroid cells. In terms of lymphoid output, 5.5% of cord blood colonies contained NK cells (the only lymphoid cell type to reliably read out in the assay used). In contrast NK cells were very rare in the adult colonies (found in < 0.05% of BM-derived colonies), supporting the findings of murine studies that there is loss of lymphoid output with ageing[71].

When looking at BM-HSC derived colonies (**Fig. 3.7b**), there was no striking change in colony phenotype with age. However, slightly more erythroid-containing colonies were observed in the older individuals, perhaps suggesting a more 'progenitor-like' phenotype for aged HSCs. In addition, more granulocyte predominant than monocyte predominant colonies were found in the individuals aged > 70. However, more work would need to be performed to provide robust conclusions on these points.

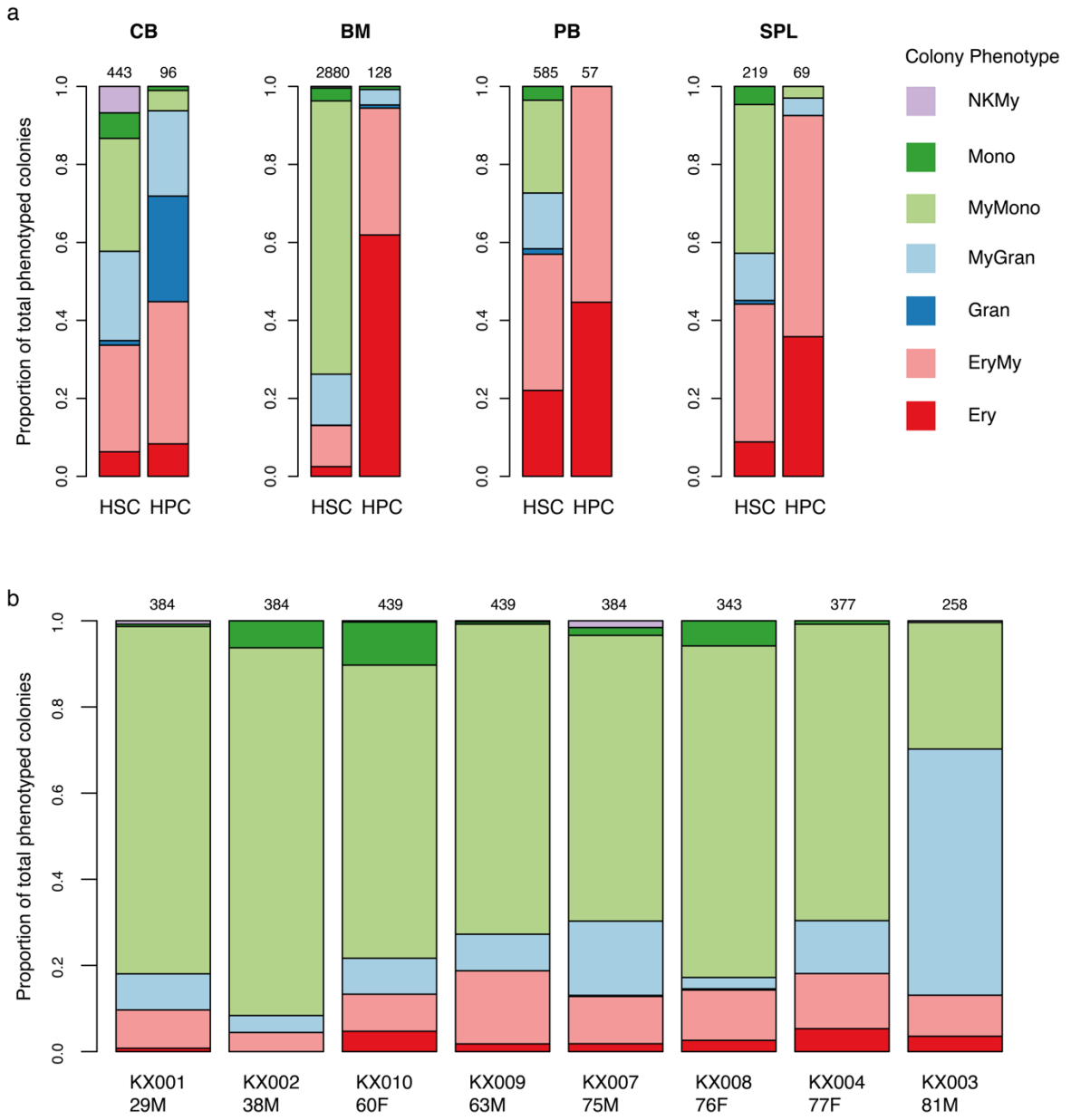


Fig. 3.7| Colony phenotypes. a, Colony phenotypes compared by sample type and seeding cell type. Ery = erythroid cells only; EryMy = erythroid cells and myeloid cells (any combination of monocytes and granulocytes); Gran = granuloctes only; MyGran = granulocytes and monocytes (granulocytes predominant); Mono = monocytes only; MyMono = monocytes and granulocytes (monocytes predominant); NKMy = NK cell and myeloid cells (any combination of monocytes and granulocytes); NKMy = NK cell and myeloid cells (any combination of monocytes and granulocytes). **b**, Phenotypes compared by donor for BM HSC-derived colonies only. Barplots coloured by phenotype as above.

Summary lineage output

1. HSC-derived colonies are significantly larger then HPC-derived colonies.
2. BM-derived colonies are significantly larger than CB, PB and SPL-derived colonies.

3. Significant decrease in BM HSC colony size with age, but linear regression contained only two individuals aged <55.
4. More striking decrease in HPC colony size as observed microscopically on manual assessment of colony size and when comparing CB HPC colony size with BM HPC colony size (only donors KX009 and KX010 had BM HPC immunophenotypic HPC data available).
5. There was no association between colony size and telomere length at any age.
6. Colony phenotype differs by seeding cell type, with HSC derived colonies containing predominantly monocytes and granulocytes, compared to a higher proportion of erythroid cells in HPC derived colonies.
7. NK cell output is almost absent in adult derived HSCs and completely absent from all HPCs.
8. BM HSC colonies show no striking change in colony phenotypes with age.

3.7 Population structure

Analysis of population structure using phylogenies created from a random sample of individuals in a population can provide information on a number of useful population level parameters. Firstly, in the absence of obvious positive selection and if generation time is constant, the pattern (or density) of historic branching events (so-called 'coalescences') provides information on historic population size changes. When considering HSC populations, the generation time is the time between successive self-renewal divisions.

Secondly, phylogenies can provide evidence for positive selection, particularly when the most recent common ancestor of a large clade contains a known driver mutation. In this context phylogenies can be used to time acquisition of driver mutations and estimate the fitness effects they confer. A 'clade' can be defined as a group of organisms descended from a single common ancestor. In the context of somatic cells, this represents a clone, and its size can be estimated from the fraction of total colonies derived from that ancestor. For this work, an 'expanded clade' is defined as a post-natal ancestral lineage, whose descendants contributed >1% of colonies at the time of sampling.

Overview of phylogeny generation

The phylogenies in **Figure 3.8** and **Figure 3.9** depict the lineage relationships between ancestors of the stem and progenitor cells sequenced in each adult individual in our cohort.

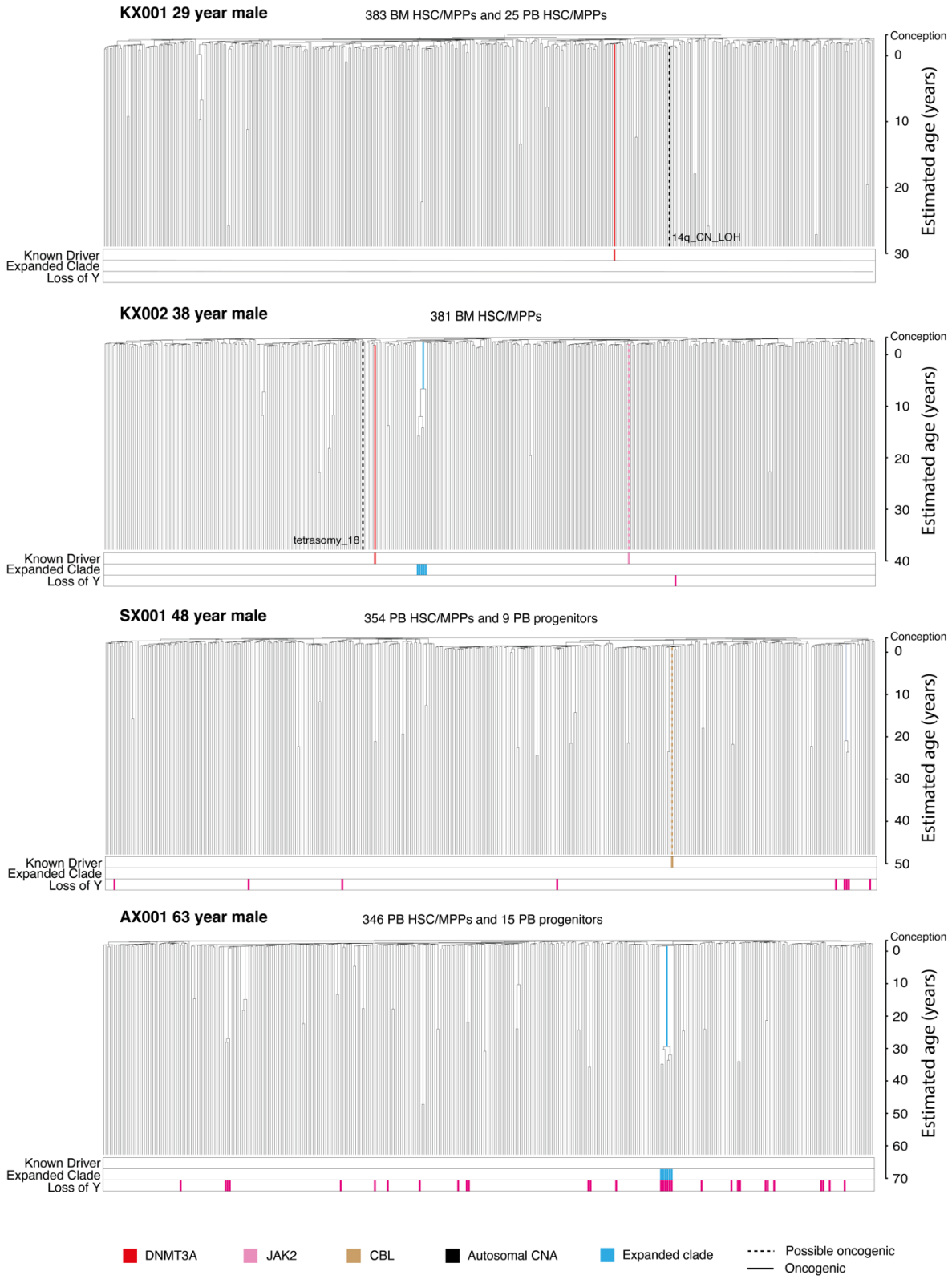


Fig. 3.8| HSPC phylogenies for the four youngest adult donors. Phylogenies were constructed using shared mutation data and the algorithm *MPBoot* (**Chapter 2**). Branch lengths are proportional to the number of mutations assigned to the branch – terminal branches have been corrected for sequence coverage, and overall root-to-tip branch lengths normalised to the same total length (because all colonies were collected from a single timepoint). The y axis is scaled to chronological time using the 'molecular clock' of somatic mutation rate, with age 0 (representing birth) set at 55

mutations (as estimated from our cord blood colonies). Each tip on a phylogeny represents a single colony, with the respective numbers of colonies of each cell and tissue type recorded at the top. Onto these trees, we have layered clone and colony-specific phenotypic information. We have highlighted branches on which we have identified known oncogenic drivers (solid line) and possible oncogenic drivers (dashed line) in one of 17 clonal haematopoiesis genes (**Table S4**), coloured by gene. Branches with autosomal copy number alterations are highlighted with a black dashed line. A heatmap at the bottom of each phylogeny highlights colonies from 'known driver' clades in red, 'expanded clades' (defined as those with a clonal fraction > 1%) in blue and colonies with loss of the Y chromosome in pink (males only). BM, bone marrow; PB, peripheral blood; CNA, copy number alteration; CN_LOH, copy-neutral loss of heterozygosity.

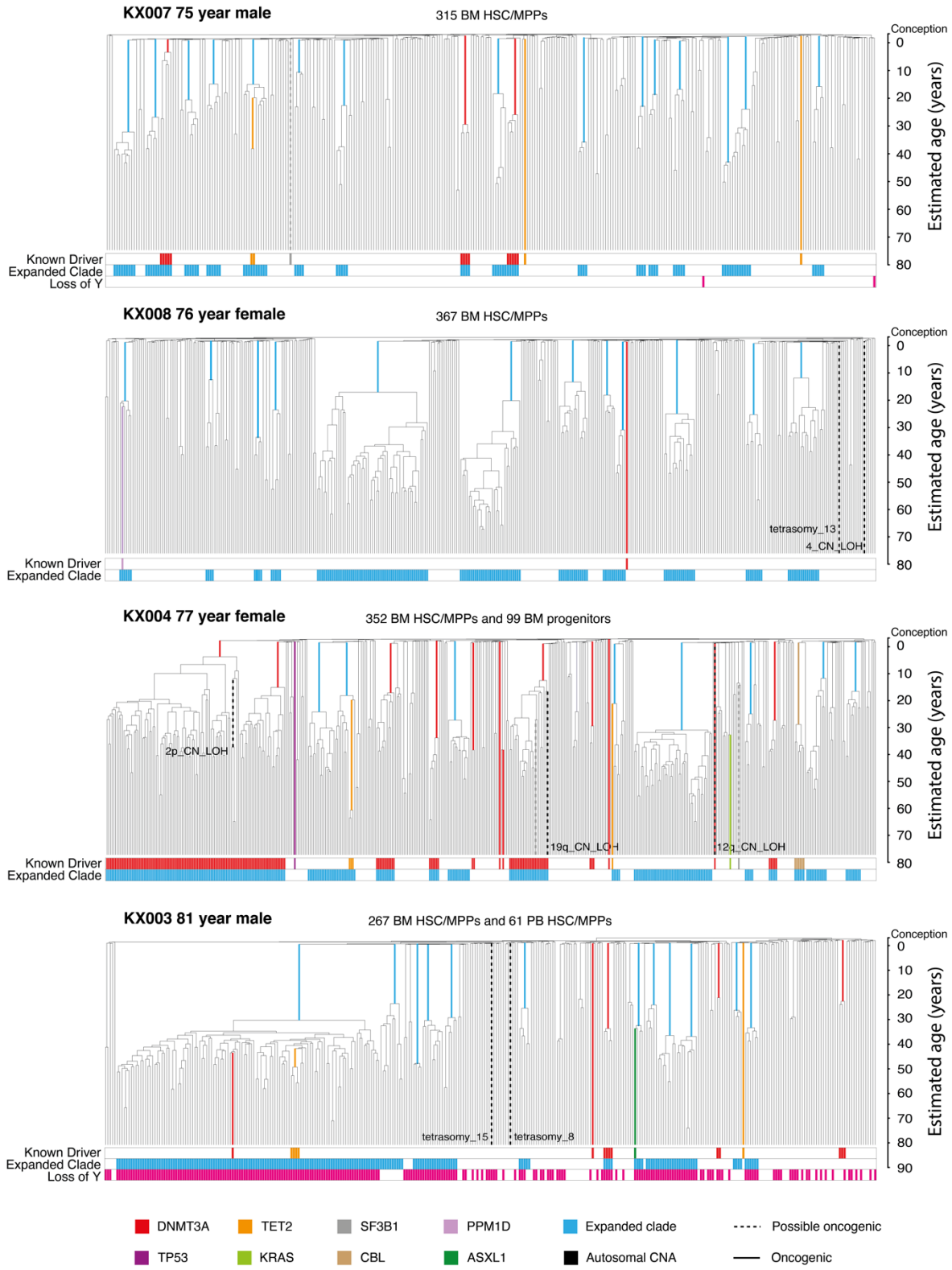


Fig. 3.9|HSPC phylogenies for the four elderly adult donors. Phylogenetic trees were constructed and presented as described for **Fig. 3.8**.

The key steps to generate the phylogenies shown in **Figures 3.8 and 3.9** are as follows:

1. Generate a 'genotype matrix' of mutation calls for every colony within a donor – the experimental approach used, based on whole genome sequencing of single-cell-derived colonies, generates consistent and even coverage across the genome, leading to very few missing values within this matrix (ranging from 0.005 – 0.034 of mutated sites in a given colony across different donors within our cohort). This generates a high degree of accuracy in the constructed trees.
2. Reconstruct phylogenetic trees from the genotype matrix – This is a standard and well-studied problem in phylogenetics. The low fraction of the genome that is mutated in a given colony (<1/million bases) coupled with the highly complete genotype matrix mean that different phylogenetics methods produce reassuringly concordant trees. The *MPBoot* algorithm was used for the tree reconstruction, as it proved both accurate and computationally efficient for our dataset.
3. Correct terminal branch lengths for sensitivity to detect mutations in each colony – The trees generated in the previous step have branch lengths proportional to the number of mutations assigned to each branch. For the terminal branches, which contain mutations unique to that colony, variable sequencing depth can underestimate the true numbers of unique mutations, so these branch lengths were corrected for the estimated sensitivity to detect mutations based on genome coverage.
4. Make phylogenetic trees ultrametric – After step 3, there is little more than Poisson variation in corrected mutation burden among colonies from a given donor. Since these colonies all derived from the same timepoint, branch lengths can be normalised to have the same overall distance from root to tip (known as an ultrametric tree). We used an 'iteratively reweighted means' algorithm for this purpose.
5. Scale trees to chronological age – Since mutation rate is constant across the human lifespan, we can use it as a 'molecular clock' to linearly scale the ultrametric tree to chronological age.
6. Overlay phenotypic and genotypic information on the tree – The tip of each branch in the resulting phylogenetic tree represents a specific colony in the dataset, meaning that phenotypic information about each colony can be depicted underneath its terminal branch (the coloured stripes along the bottom of **Figures 3.8 and 3.9**). Furthermore, every mutation in the dataset is confidently assigned to a specific branch in the phylogenetic

tree. This means that branches on which specific genetic events occurred can be highlighted (such as *DNMT3A* or other driver mutations).

More detailed information on the steps described above is provided in **Chapter 2**.

Changes in population structure over the human lifespan

HSC phylogenies for the 4 adults aged < 65 have few coalescences and a paucity of clonal expansions, showing that haematopoiesis in healthy young and middle-aged individuals is typically highly polyclonal. Despite sequencing 361-408 colonies per individual, at most a single expanded clade was identified in each individual. The two expanded clades that were observed contributed < 2% of all haematopoiesis. Only 4 known or possible driver mutations were identified in any one of the top 17 clonal haematopoiesis genes (listed in **Table 3.4** below). This gene list was used to define which mutations to depict on the phylogenies as these genes were found to be under significant positive selection in a recent large longitudinal clonal haematopoiesis dataset[133].

Table 3.4: Top 17 clonal haematopoiesis genes

DNMT3A	TET2	ASXL1	SRSF2	U2AF1	PPM1D
SF3B1	TP53	JAK2	CBL	KRAS	GNB1
CTCF	BRCC3	IDH1	PTPN11	IDH2	

In contrast, the phylogenies of the 4 adults aged > 70 showed a markedly oligoclonal pattern of relationships between the sampled cells. Each elderly individual had between 12-18 independent expanded clades established between birth and the age of 40. These expanded clades each comprised between 1% and 34% of colonies sequenced, most in the 1-3% range. Together, these clades made up a significant proportion of the HSC/MPP population in our elderly research subjects in whom 32-64% of all colonies sequenced derived from expanded clades.

Surprisingly, only a minority of clonal expansions in the elderly individuals carried known driver mutations. Considering the top 17 clonal haematopoiesis genes listed above, mutations in *DNMT3A*, *TET2* and *CBL* were identified as causal drivers in 10/58 expanded clades (those with a clonal fraction > 1%). Only an additional 3 causal mutations were identified if we extended to a wider set of 92 genes implicated in myeloid neoplasms[106] (**Table S2**). Thus, known clonal haematopoiesis mutations were only able to explain 22% of the observed clonal expansions, leaving 45 expansions (78%) unexplained.

Population dynamics in young adults

As discussed in **Chapter 1** and **Appendix 1**, the study of phylodynamics has shown that the frequency of coalescences in phylogenetic trees in a neutrally evolving, well-mixed population of somatic cells is primarily determined by the product of population size and time between symmetric self-renewal cell divisions ($N\tau$) – both smaller populations and more frequent symmetric self-renewal divisions increase the density of coalescences. In young adults, where clonal selection has had minimal impact on the HSC phylogenetic tree structure, we can utilise phylodynamic approaches to estimate the trajectory of HSC population size dynamics[168]. **Figure 3.10a** shows population trajectories generated by *phylodyn*[169] for the four adults aged < 65. *Phylodyn* uses the density of coalescences to determine $N\tau$, with each bump in the black trajectory representing a coalescence event in the phylogeny from which it is derived. The trajectories show a rapid increase in $N\tau$ during *in utero* development, reaching the level seen in young adults close to the time of birth. There is some evidence that $N\tau$ may be higher in young childhood (0-7 years) before stabilising to adult levels after this time. For all four young adults *phylodyn* plateaus at an $N\tau$ of around 100,000 HSC-years in late childhood / young adult life.

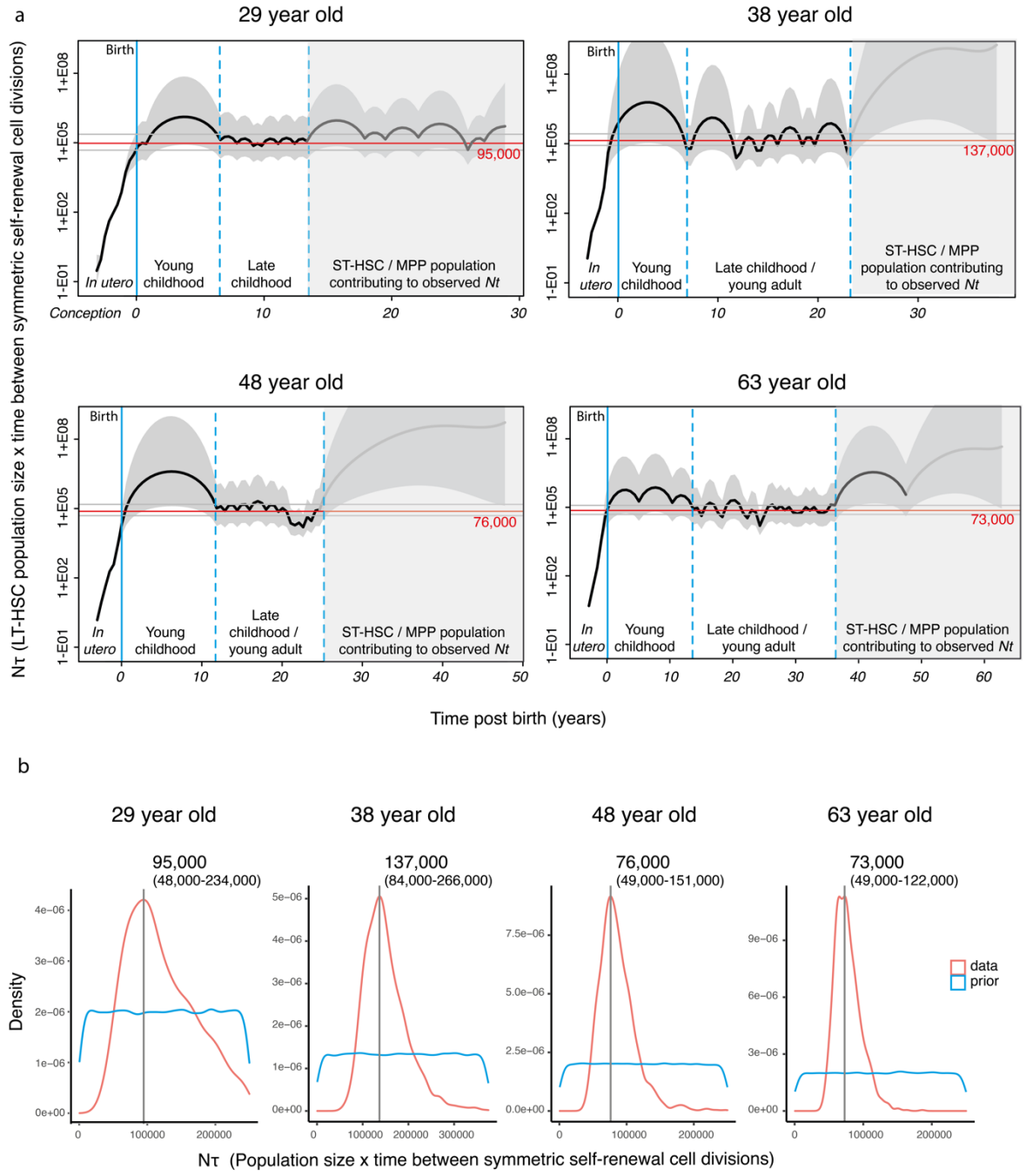


Fig. 3.10| Estimating $N\tau$ in the human LT-HSC compartment. a, *Phylodyn* plots illustrating the trajectory of $N\tau$ for human LT-HSCs in the four adult donors aged <65. The black line represents the trajectory of LT-HSC $N\tau$, with the shaded grey area on either side representing the 95% credibility interval. The solid blue line is the time of birth. The dashed blue lines enclose the region of time in each individual where the trajectory is at the late childhood/young adult level. The shaded region of the plots represents the period of time prior to sampling over which it is likely that ST-HSC/MPPs are contributing to the observed $N\tau$. The trajectory line is shaded dark grey in the time period where coalescent events are occurring and the trajectory likely represents the combined LT-HSC and ST-HSC/MPP $N\tau$. The trajectory line is shaded light grey where there is a complete absence of coalescent events and the estimates are highly inaccurate. The red line shows the Bayesian (maximum posterior density) estimate of $N\tau$. **b,** Results from approximate Bayesian inference of

population size over the first (non-shaded) part of life for each individual. The blue line represents the prior density of $N\tau$ and the red line represents the posterior density. The vertical grey line is the peak $N\tau$ for each donor. The peak $N\tau$ with 95% confidence limits is written at the top of each plot.

Approximate Bayesian Computation (ABC) was used as an orthogonal approach to estimating $N\tau$ in young adulthood from the phylogenetic trees of adults aged < 65. ABC is an approach rooted in Bayesian statistics that can be used to estimate the posterior distributions of model parameters. The principle of the approach is to create a model of how the observed data of interest is generated. In this case a model of the full HSC population was used, from which a phylogeny comparable to our real phylogenies could be generated. HSC population size can then be varied between runs of the model (or simulations), and summary statistics can be used to compare the simulated phylogenies with the real phylogenies to identify a posterior distribution of HSC population size that best fits the observed data.

In this case, an HSC population simulator (*rsimpop*) was used to simulated 100,000 population trajectories for each individual with a population size (N) in the first 2-3 decades of life selected from a uniform distribution between 1000 and 250,000 (KX001, SX001, AX001) or 400,000 (KX002) and symmetric self-renewal division rate set at 1 per year. The summary statistics used for the ABC modelling were time-weighted mean number of lineages calculated at 3 evenly spaced time points through the phylogeny. These statistics were used to select the top 1% of phylogenies that best fit the observed phylogeny for each individual. **Figure 3.10b** shows the posterior distribution of $N\tau$ for the top 1% of phylogenies. These estimates are also overlaid in red on the *phylodyn* plots, showing high concordance between the two approaches, with $N\tau$ calculated for all four adults being close to 100,000 (95%CI in the range 50,000 – 250,000) (**Fig. 3.10a**).

Estimating HSC population size in young adults

The values for $N\tau$ estimated above were consistent across all four individuals, and in keeping with published estimates[86,143]. While the estimates of $N\tau$ across these individuals are highly concordant, it is likely that HSC population size will vary between individuals, mostly likely following a normal distribution. A number of factors including germline modifiers, body size,

sex and co-morbidities (such as chronic inflammation), may play a role in determining HSC population size in a given individual.

In order to put bounds on *N* (the HSC population size), we needed an estimate of τ, the time in years between symmetric HSC self-renewal divisions. Previous estimates of τ from the literature range between 0.6 and 6 years[42,86,88]. Our telomere data, together with published estimates of rates of telomere loss per division and the proportion of HSC divisions that are symmetric, also allows us to independently estimate τ. We have shown in adult life that HSC/MPP telomeres shorten at a rate of 30 bp/year. In addition HSC telomeres are known to lose 30-100bp per cell division[58] and two recent studies in mice have shown that symmetric divisions predominate within the HSC pool, making up 80-100% of all HSC divisions[170,171].

Using the estimates above in a Monte Carlo simulation approach, we sampled from the distributions of each variable 500,000 times, calculating the value of *N* for each set of randomly sampled variables. This was done using the following distributions: telomeric shortening rate per division: uniform(minimum = 30, maximum = 100); symmetric division rate: uniform(minimum = 0.8, maximum = 1.0); $N\tau$: uniform(minimum = 50,000, maximum = 250,000); average telomere loss per year: uniform(minimum = 30, maximum = 40). Using this approach, these data are most consistent with adult haematopoiesis being maintained by a population of 20,000-200,000 long-term HSCs.

Evidence for a long-lived ST-HSC/MPP compartment

In both young and old individuals there was a paucity of coalescent events towards the tips of the phylogenetic trees compared to earlier ages (**Figs. 3.8 and 3.9**). This change occurred consistently around 10-15 years of molecular time prior to sampling irrespective of the age of the individual. The change in density of coalescent events over this period can also be visualised as an increase in $N\tau$ in the *phylodyn* trajectories in **Figures 3.10 and 3.12**. The most plausible explanation for the observed pattern is that it is caused by the existence of a large, short-term HSC/MPP compartment in humans that is able to self-renew for 10-15 years. Due to the short-term nature of its contribution, the effects of this compartment would only be visible towards the tips of the phylogenetic trees, with earlier time-points more accurately reflecting dynamics in the LT-HSC compartment. The substantially larger size of a combined

long-term and short-term HSCs/MPP population would explain the reduction in coalescence density in the 10-15 years prior to the time of sampling**. Figure 3.11** shows simulated phylogenies with increased population sizes towards the time of sampling, illustrating the population size changes that could cause the pattern we observe. Work in mice has shown that their equivalent ST-HSC/MPP compartment is able to sustain steady state haematopoiesis without significant input from the long-term HSC compartment for up to a year, which would fit with our hypothesis of an analogous long-lived ST-HSC/MPP compartment in humans[172].

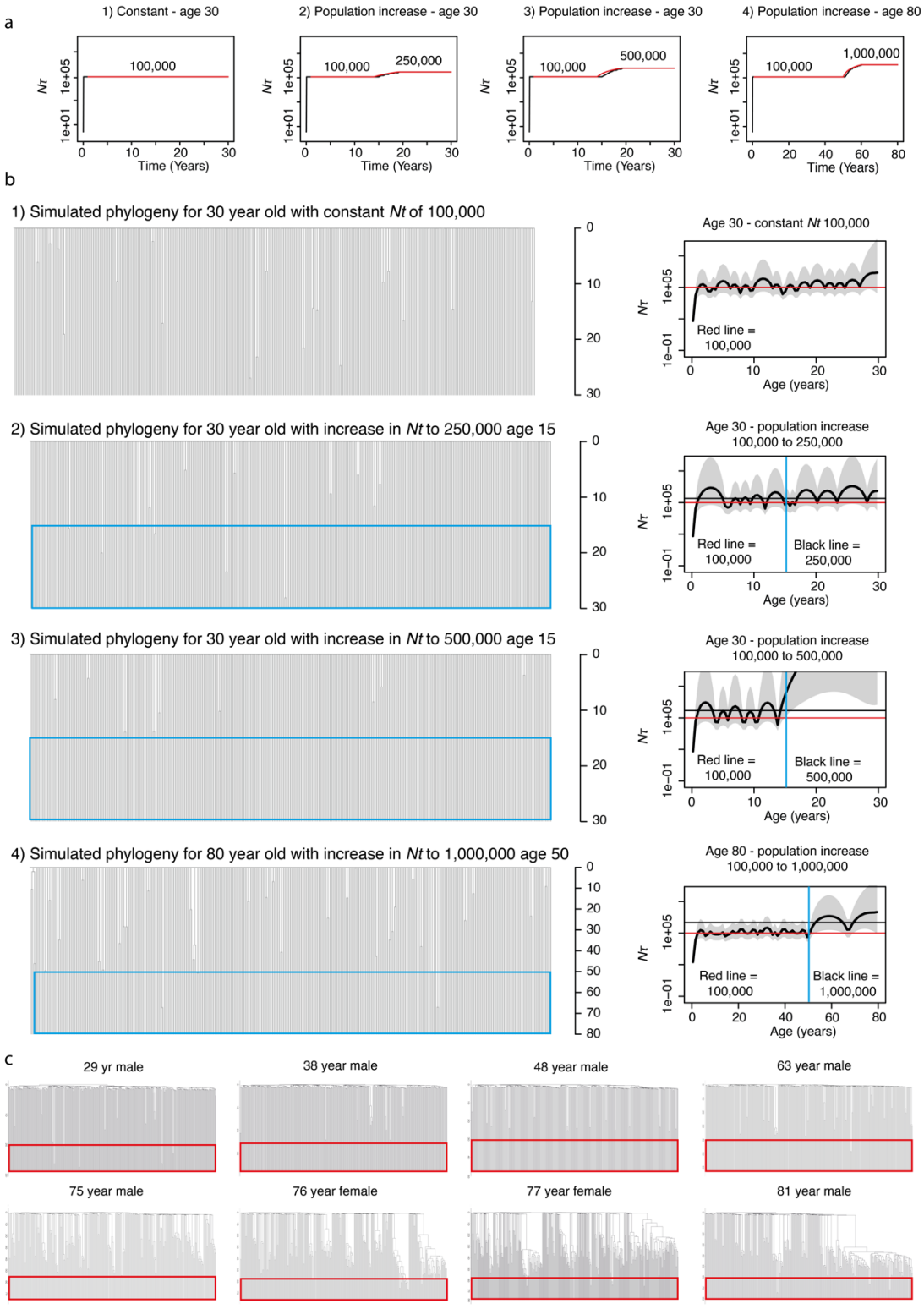


Fig. 3.11| Interpretation of young adult HSPC phylogenies. a, Trajectories of $N\tau$ used as input to *rsimpop* for the simulations to create phylogenies in b. Note the Y axis depicting $N\tau$ is on a log scale. **b,** Phylogenies created by randomly sampling 380 cells from the final full simulated population of between 100,000 cells (Phylogeny 1) and 1,000,000 cells (Phylogeny 4). Phylogenies 1 to 3 are derived from simulations of the HSC population in a 30-year-old, while phylogeny 4 is derived from a simulation of the HSC population in an 80-year-old. Each simulation has an initial $N\tau$ of 100. In all

cases $N\tau$ is the same as the population size (N) as the generation time (τ) in all simulations is fixed at 1. The blue boxes indicate the period of time in which the population size is increased. The *phylodyn* trajectories to the right of each simulated phylogeny use the pattern of coalescent events to recover the input trajectories for $N\tau$. The blue line marks the time of change in $N\tau$. In all cases the initial part of the trajectory is able to correctly estimate $N\tau$ at 100,000. However, in Phylogeny 3 where there is a complete absence of coalescent events once the population size is increased, *phylodyn* loses resolution and wildly overestimates the value of $N\tau$. **c,** Real trees with red boxes highlighting the last 10-20 years prior to sampling, where the relative number of coalescent events is decreased (meaning the estimated $N\tau$ is larger).

Population dynamics in old age

The presence of expanded clades with known driver mutations in the elderly phylogenies provides evidence for a role for positive selection in shaping the pattern of coalescent events in these individuals. Evidence of positive selection in a phylogeny invalidates use of the phylodynamic approach to inferring population dynamics that we were able to use for the adults aged <65. Nevertheless, the *phylodyn* population trajectories for the elderly individuals are shown in **Figure 3.12**, which illustrate that if these elderly HSC populations had undergone only neutral evolution, the trajectories would be most compatible with a population 'bottleneck' in mid-life. *Phylodyn* does not take into account the distribution of coalescent events across the clades in the phylogeny, only the number of coalescent events within a given time window. In addition, there is good evidence that the HSC populations in the elderly have undergone at least some positive selection (for example the presence of expanded *DNMT3A* mutated clades), invalidating the use of *phylodyn*, and its interpretation of the pattern of coalescent events through time as population bottlenecks.

In **Chapter 4** additional analyses are presented which show that mid-life population bottlenecks are unable to explain the pattern of coalescences observed in the elderly phylogenies. While simulations incorporating population bottlenecks could generate phylogenies with the same number of coalescent events in midlife, they could not recapitulate the asymmetry of clade sizes found in the real phylogenies. We therefore do not think that significant population bottlenecks are occurring in HSC populations in midlife in normal individuals.

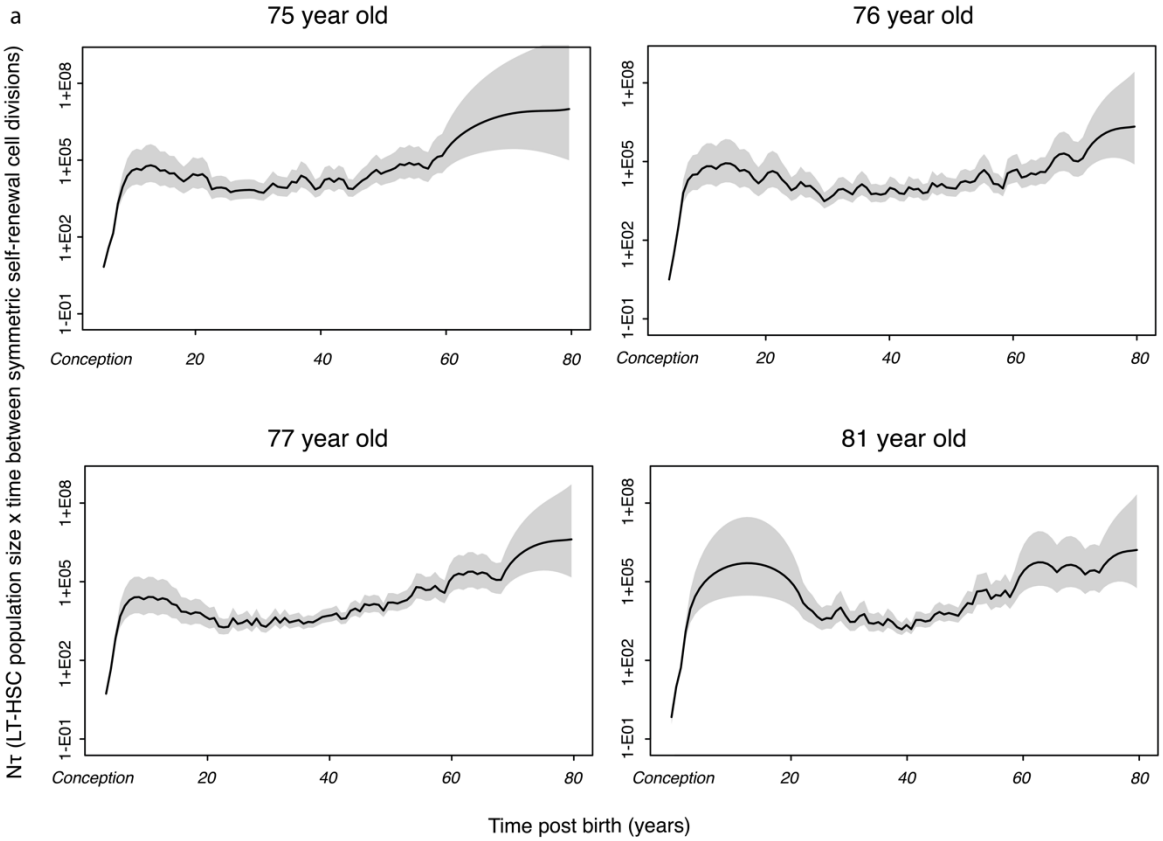


Fig. 3.12| Phylodyn plots for elderly individuals. *Phylodyn* plots illustrating the trajectory of $N\tau$ for human LT-HSCs in the four adult donors aged >75 if the pattern of coalescent events in their respective phylogenies was not confounded by the presence of positive selection. The black line represents the trajectory of LT-HSC $N\tau$, with the shaded grey area on either side representing the 95% credibility interval. However, given that we have identified at least some clonally expanded driver mutation containing clades, these population trajectories cannot be considered reliable.

Figure 3.13 shows two simulated phylogenies for an 80 year-old. The first depicts a neutrally evolving HSC population that remains static in size over life, and the second depicts a neutrally evolving population with a population bottleneck in mid-life.

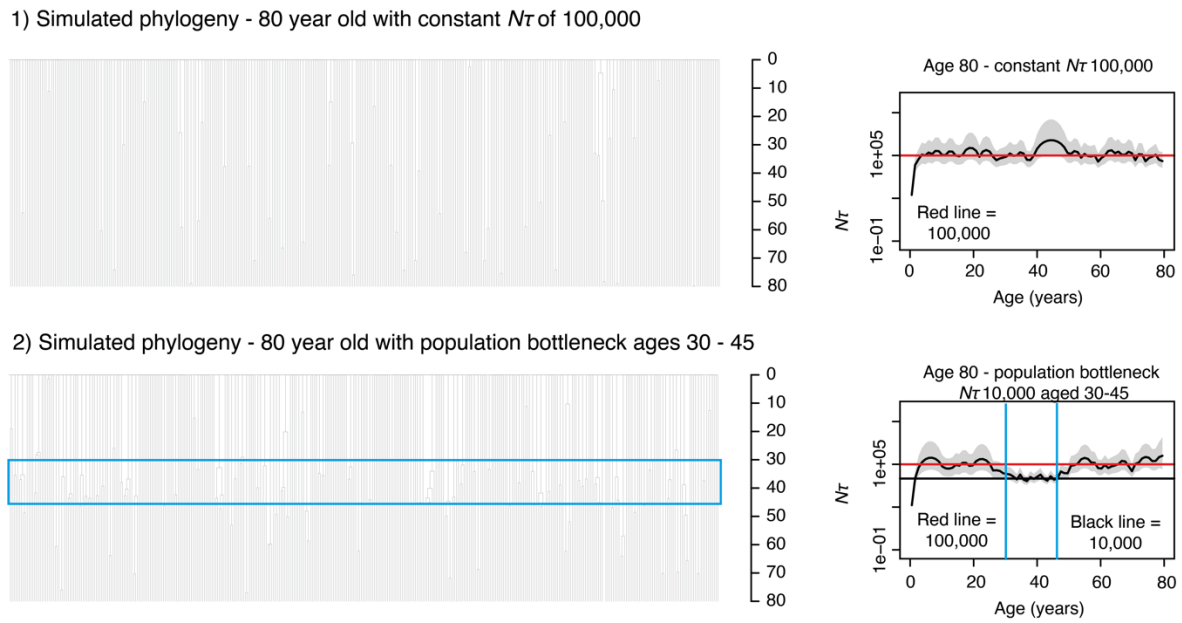


Fig. 3.13| Interpretation of elderly adult HSPC phylogenies. Phylogenies created by randomly sampling 380 cells from the final full simulated population. As with the previous simulations, $N\tau$ ~ population size because the time between symmetric self-renewal divisions is set at 1 year. In both simulated phyologenies the final population is 100,000 cells in size, but Phylogeny 2 has been created from a population that underwent a bottleneck in size to and $N\tau$ of 10,000 cells between the ages of 35 and 50. This period of time over which the population size was reduced can be visualised in the blue box which highlights the increased density of coalescent events in this time block. The phylodyn trajectories are able to accurately recover information on changes in $N\tau$ of the HSC population over time.

At the outset of the project one aim was to characterise LT-HSC population size changes with ageing. However, the presence of both positive selection and a contribution of short-term HSC/MPPs in the 10-15 years prior to sampling makes it impossible to determine how the LT-HSC population may change in size in later life. Nevertheless, our data does suggest that the $N\tau$ of the combined LT-HSC and ST-HSC/MPP compartment does not decrease significantly in old age, and would support previous observations that the combined compartment may increase with age as has been well documented in mice[173] and to a lesser extent humans[174].

Summary population structure

1. Haematopoiesis in healthy, young adults is highly polyclonal, with few known driver mutations detectable.
2. Haematopoiesis in adults aged > 70 is universally oligoclonal with between 32-64% of HSC/MPPs sampled deriving from expanded clades.

3. Known driver mutations explain 13/58 (22%) of the expanded clades identified in the elderly individuals.
4. LT-HSC population size was estimated at between 20,000 and 200,000, using $N\tau$ estimated from the young adult phylogenies (50,0000 to 250,000) and τ derived from the estimate of HSC telomere loss per year.
5. A paucity of coalescences in the 10-15 years of molecular time prior to sampling provides evidence for a large population of ST-HSC/MPPs that are able to self-renew over this time period.
6. There is no evidence for a decline in the size of the combined LT- and ST-HSC/MPP population in old age.

3.8 Age-related change in HSCs summary

In this chapter a survey of cell intrinsic age-related changes in human HSCs has been undertaken. Key novel findings have added considerably to what was previously known in a number of areas. Firstly, there is no accelerated somatic mutation accumulation in old HSCs. Secondly, the observed distribution of HSC telomere lengths provide evidence for a more dormant population of HSCs that is lost in older age. Thirdly, the combined population of LT and ST/HSCs remains stable or increases in size in old age. Finally, there is a striking change from polyclonal haematopoiesis in young and middle-aged adults to markedly oligoclonal haematopoiesis in the elderly. The mechanisms underlying this profound change in the clonal structure of HSC populations with age will be explored in **Chapter 4**.

Chapter 4: Mechanisms underlying change in HSC population structure with age

4.1 Introduction

In **Chapter 3**, HSC phylogenies created from individuals of different ages revealed a dramatic change in HSC population structure occurring between the ages of 65 and 75. Over the age of 70, 32-64% of haematopoiesis is derived from clonally expanded clades, compared to less than 2% in all four younger individuals aged < 65 (**Fig. 4.1a**). This is well-illustrated by measures of clonal diversity, such as the Shannon Index which, when applied to our phylogeny data, show a precipitous decline after the age of 70 (**Fig. 4.1b**). The timing of the change from polyclonal to oligoclonal haematopoiesis largely mirrors the time of onset of age-related loss of function in the haematopoietic system.

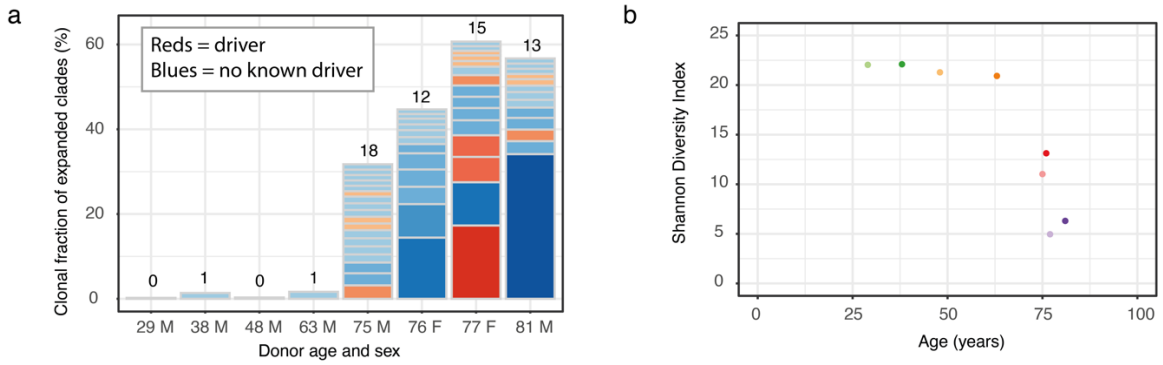


Fig. 4.1| Oligoclonality in the HSC/MPP compartment of normal individuals. a, Stacked barplot to show number and size of clades with a clonal fraction >1% per individual. Shades of blue denote clones with no known driver and shades of red denote those with a known myeloid gene driver. The numbers above each stack denote the total number of expanded clades in the individual. **b,** Shannon diversity index calculated for each phylogeny. The Shannon diversity index is a mathematical measure of species diversity that takes into account both the abundance and evenness of the species present. Here the number of species is the number of lineages present at 100 mutations of molecular time (equivalent to the first few years after birth) and abundance is the number of samples within the clade derived from that lineage.

The mechanisms underlying the profound change in HSC population structure with age are the focus of this chapter. There are a number of possible hypotheses that could explain the observed population structure changes. Firstly, age-related changes in HSC population size,

similar to those depicted in the inferred elderly population trajectories in **Figure 3.12**, could explain the observed pattern of expanded clades in the real elderly phylogenies. Secondly, changes in spatial patterning of HSCs in the bone marrow with age could be implicated in the change in population structure. Thirdly, positive selection could be acting on variation in cell intrinsic factors, including unknown driver mutations and epigenetic changes, to drive clonal expansions. The chapter concludes with analysis of the timing and fitness effects of driver mutations as well as investigating possible functional consequences of the accumulation of driver mutations in the largest clades.

Key questions to be addressed in this chapter

1. Could HSC population size changes over life explain the observed pattern of expanded clades in the real elderly phylogenies?
2. Is there evidence for pervasive positive selection on previously unknown drivers?
3. Could selection on previously unknown driver mutations (or cell intrinsic factors) explain the observed abrupt change in HSC population structure with age?
4. When in life are driver mutations that clonally expand by old age being acquired and what fitness effects do they confer?
5. Can any functional effects of driver mutations be determined?

These questions were investigated using a number of approaches. Firstly, Approximate Bayesian Computation modelling[130,131] was performed using an HSC population simulator (*rsimpop*) to model the effect of both population size changes and driver mutation acquisition. Secondly, a genetic analysis (*dNdScv*)[175] was used to identify positive selection in the dataset, agnostic to which genes may be under selection. Thirdly, the phylogenies were used to time driver mutation acquisition and calculate fitness effects of drivers contributing to expanded clades. Finally, immunophenotyping data was interrogated for information on clade specific functional effects of known and unknown driver mutations.

4.2 Population size modelling

Approach to population size modelling

In an exactly analogous approach to the modelling used to derive estimates of $N\tau$ in **Chapter 3**, ABC population modelling was used to investigate whether changes in population size with age could recreate the population structures observed in the elderly phylogenies. The *phylodyn* trajectories for each individual (**Figs. 3.10 and 3.12**) were used to identify the timing of midlife and late-life fold-change in $N\tau$ most compatible with the pattern of observed coalescences, which were used to inform the population trajectory models. The most likely pattern of population size change that could explain the observed elderly HSC phylogenies was a population 'bottleneck' occurring in midlife. However, as already discussed at the end of **Chapter 3**, the presence of significant positive selection in a population invalidates the use of *phylodyn* and its population size trajectory predictions.

Mutation burden estimates were used to inform the mutation rate, which was set at 15 mutations per year with an additional 1 mutation for every cell division. Telomere attrition rates were used to inform the time between symmetric cell divisions, which was set at 1 year (after the initial population growth phase in early life). A large sample of simulated HSC populations was generated for each individual, in which initial population size (*N*) and fold population size changes in midlife and late-life were varied. Summary statistics (the number of lineages in the phylogeny through time) were used to compare the phylogenies drawn from the simulated populations, with those from the real individuals. See **Figure 4.2** and **Chapter 2** for more details about the modelling approach, parameters and summary statistics used.

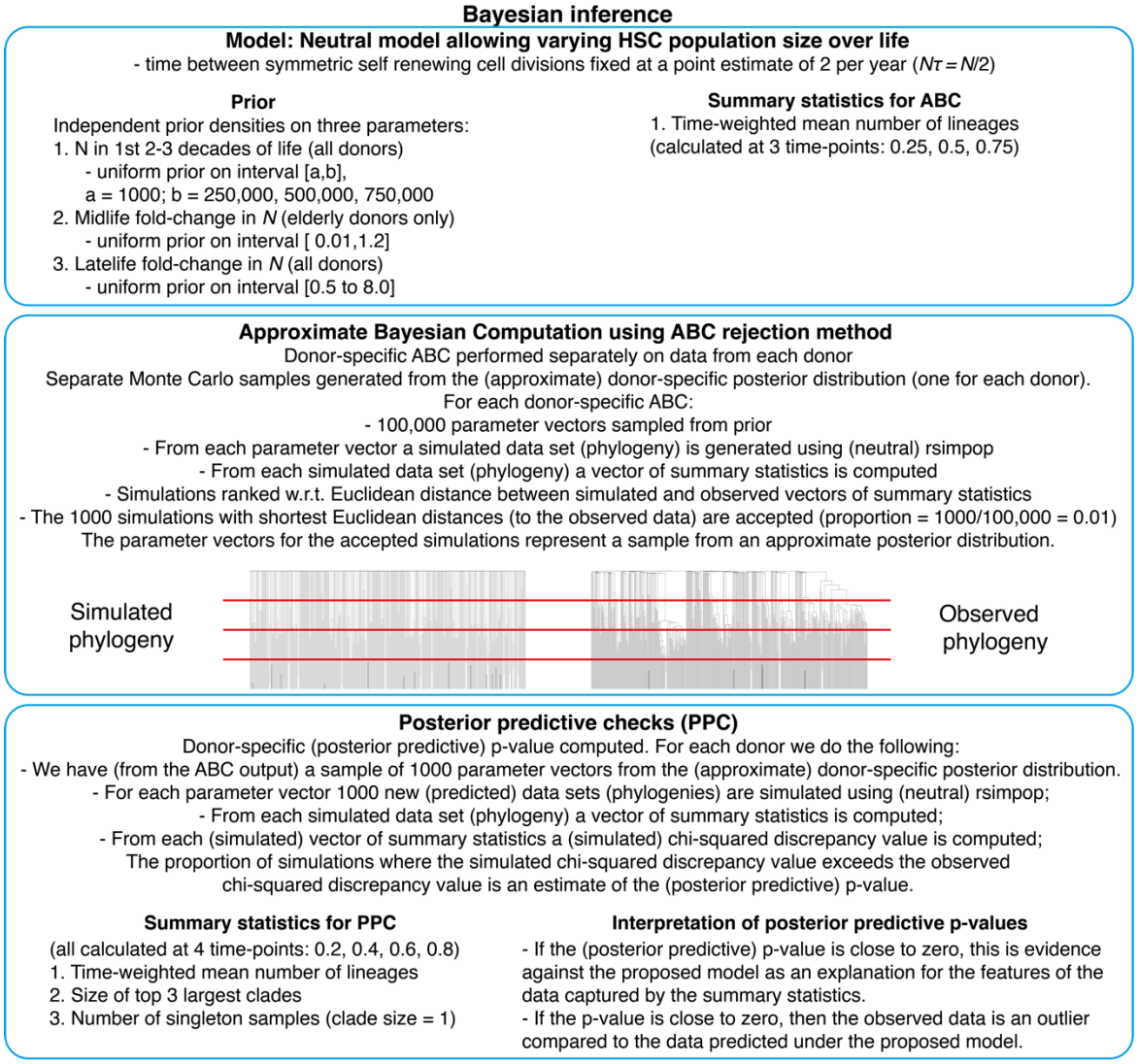


Fig. 4.2| Modelling HSC populations incorporating only changes in $N\tau$, without positive selection. Overview of modelling approach used to estimate $N\tau$ alone in the young adult individuals and to investigate whether changes in $N\tau$ could explain the observed clade size distribution in the elderly adult individuals. These simulations were run using a neutral model (that is, no acquisition of driver mutations), with $N\tau$ being the only parameter to change over time. For the young adult individuals $N\tau$ was estimated for two time-blocks (time before and after population increase due to ST-HSC/MPP contribution). For the elderly adult individuals $N\tau$ was estimated for three time-blocks as the *phylodyn* plots predicted a population 'bottleneck' (**Fig. 3.12**) was the most parsimonious way to recreate the observed change in coalescence density over life.

Population size modelling results

The simulations incorporating population size changes were able to accurately recapitulate the phylogenies of the four individuals aged under 65, demonstrating that the simulation approach utilised was in theory able to well match observed data for young individuals.

However, even the 1% most closely matching phylogenies from these simulations poorly replicated the phylogenies observed in the elderly (**Fig. 4.3**). While the simulations incorporating a population 'bottleneck' in midlife were able to generate phylogenies with the same number of midlife coalescences, they were unable to match the observed clade size distribution. The observed elderly phylogenies harbour both a few large clades but also numerous 'singleton' branches. In contrast, the simulated phylogenies generated by varying the population size over life showed a more even distribution of clade sizes, consisting entirely of a few large clades if the population bottleneck is very tight (1000 cells), or of evenly sized smaller clades if the population bottleneck is less restrictive (10,000 cells).

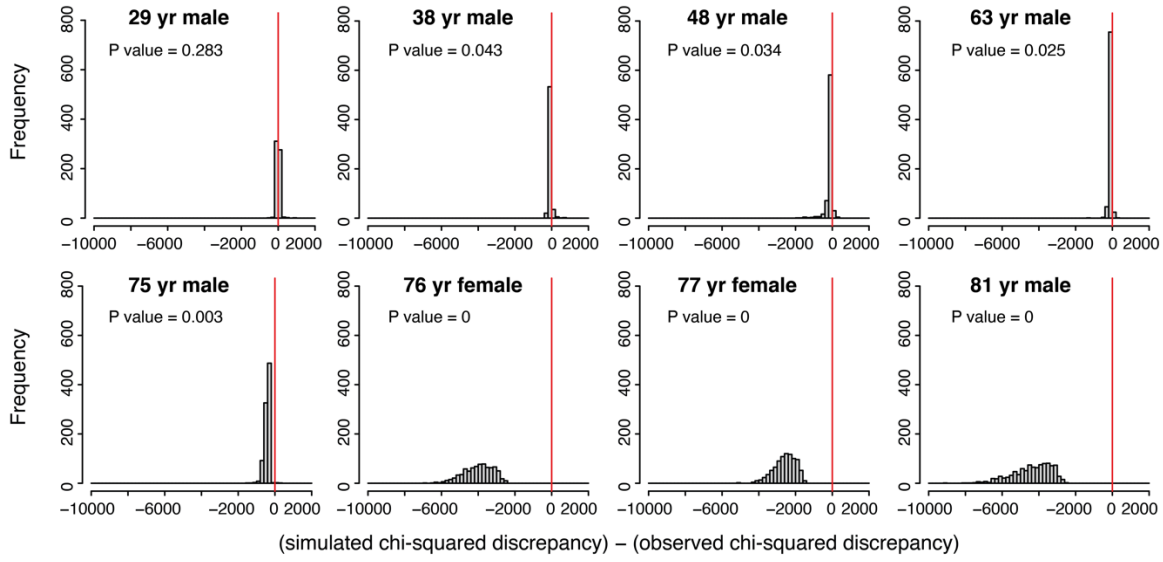


Fig. 4.3| Results of modelling HSC populations incorporating only changes in $N\tau$. Plots showing the posterior predictive distribution of the difference between the simulated chi-squared discrepancy and the observed chi-squared discrepancy, for each donor individual under a neutral model incorporating change in population size. For each donor, the posterior predictive distribution of the difference between predictive (simulated) and observed chi-squared discrepancy is represented as a histogram based on a Monte Carlo sample of 1,000,000 simulated phylogenies, drawn from the posterior predictive distribution. The proportion of simulated phylogenies which lie to the right of zero (red line) is a Monte Carlo estimate of the posterior predictive p-value (the probability that the predictive chi-squared discrepancy exceeds the observed chi-squared discrepancy under the neutral model). In the case of the four young adult individuals, the proportion of simulated phylogenies which lie to the right of 0 (red line) is close to 0.5, indicating that the simple neutral models (incorporating changes in Nτ over life) predict trees that have similar clade size distributions to our observed trees. In contrast, for the four elderly adults, the proportion of simulated phylogenies which lie to the right of 0 is very small (Less than 0.05), demonstrating

that the neutral models are, on their own, unable to recreate trees with similar clade size distributions to those observed.

The ABC modelling therefore results in rejection of population size changes with age as an explanation for the oligoclonal HSC population structures observed in the elderly. It shows that the data are not compatible with a population bottleneck in midlife. Additional support for this conclusion comes from the fact that there is no evidence for such a population bottleneck from orthogonal approaches or other datasets.

Summary population size modelling

1. Simulations incorporating population size changes with age were able to accurately recapitulate the HSC phylogenies of individuals aged < 65.
2. Simulations incorporating population size changes with age were unable to accurately recapitulate the HSC phylogenies of individuals aged > 65, allowing us to reject population size changes as the mechanism underpinning the observed change in HSC population structure in the elderly.
3. There is no support for an HSC population bottleneck in midlife explaining the observed phylogenies.

4.3 Spatial and compartmental segregation

Comparison of bone marrow and peripheral blood phylogenies

Changes in the spatial patterning of HSCs in the bone marrow could also in theory explain the observed changes in phylogeny structure with age. If HSCs become more spatially restricted in the bone marrow with ageing, then the bone marrow population from a limited number of sites would be less well mixed and contain clades of more closely related cells as observed in the elderly phylogenies.

To address the possibility of spatial segregation, for one elderly individual (KX003, 81 year male) HSC/MPPs from both BM (267 cells) and PB (61 cells) were sequenced. If spatial segregation in the bone marrow was significant, the clonal expansions evident in a BM-derived phylogeny should be much smaller, or 'diluted' down in a PB-derived phylogeny,

which would be predicted to represent an even mix of clonally expanded clades from all regions of bone marrow.

The PB HSC/MPP colonies for the 81 year male recaptured most of the expanded clades present in the BM HSC/MPP phylogeny at very similar clonal fractions (**Fig. 4.4**), suggesting that spatial segregation is not sufficient to explain the changes in the phylogenies observed with ageing. This finding is in line with previous work showing that the clonal fraction of driver mutations in patients with MDS is concordant between BM and PB samples[14].

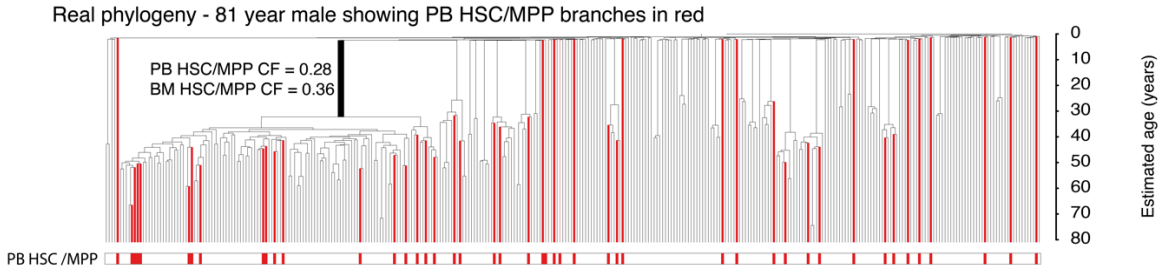


Fig. 4.4| Comparison of BM and PB derived phylogenies. Real HSC/MPP phylogeny for KX003 (81-year-male) with PB HSC/MPP terminal branches coloured red (BM HSC/MPP branches remain black). The CF of the largest clade is shown for PB and BM cells.

Comparison of HSC and HPC phylogenies

As well as investigating the possibility of spatial segregation, I was also interested in addressing whether there was any evidence for 'compartmental segregation' in the clonal expansions. The term 'compartmental segregation' refers to the situation where there are clades that are predominantly expanded either in the HSC/MPP compartment or in the HPC compartment rather than being evenly expanded in both compartments. In support of this possibility, it has previously been shown that mutations in *JAK2* commonly cause significant amplification at the level of erythroid progenitors, rather than in the HSC/MPP compartment[176].

The possibility of compartmental segregation was investigated in another of the elderly individuals (KX004, 77 year female), for whom sequencing was performed on both BM-derived HSC/MPPs (352 cells) and HPCs (99 cells). In this comparison, there were a minority

of clades whose clonal fractions differed significantly by compartment (**Fig. 4.5**). This suggests the presence of some clade-specific compartmental segregation, but not a universal bias across all clades.

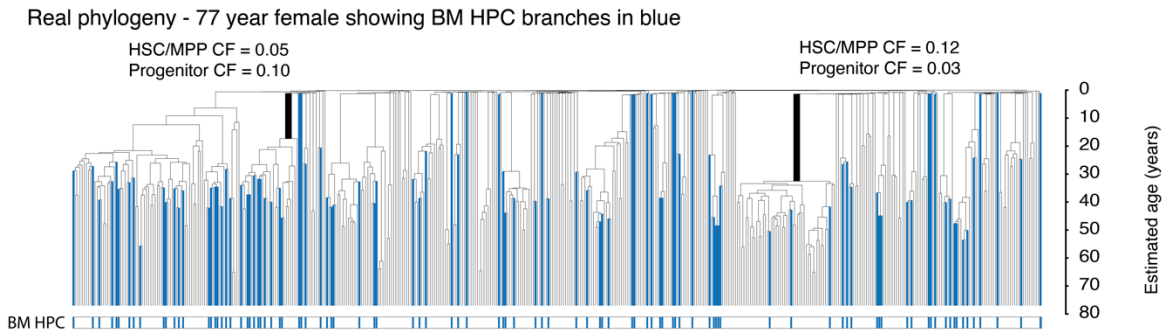


Fig. 4.5| Comparison of BM HSC and HPC derived phylogenies. Real phylogeny for KX004 (77 year female) annotated by cell type (BM HSC/MPP vs BM HPC). Two clades with differing clonal fractions of these cell types are highlighted.

Summary spatial and compartmental segregation

1. PB and BM derived HSC/MPP clonal fractions are highly concordant, suggesting there is no increase in the spatial segregation of HSCs in the bone marrow with age.
2. A minority of clades exhibit clade specific 'compartmental segregation', with a clonal fraction much higher in either HSC/MPPs or HPCs.

4.4 Pervasive positive selection

Cell autonomous variation in fitness

A prerequisite for the action of positive selection is cell-autonomous variation in fitness amongst the cells within a population. Factors that might vary within a population and provide a fitness advantage to specific clones include any somatically heritable properties. Somatic mutations are the most obvious example, but epigenetic changes, telomere lengths, and irreversible protein aggregates among other features could also provide the same basis for selection to act. In order to identify whether positive selection could underlie the pattern of coalescences in the elderly phylogenies, two approaches were used. First was a genetic approach (dN/dS) which only considers somatic mutations. The second approach involved ABC modelling of selection within the population. The ABC approach is more holistic in that it

is agnostic to the factor(s) changing cell autonomous fitness. Although in the analysis we have used the term 'driver mutations' to refer to the factor under selection.

dN/dS approach

Having excluded population size changes and spatial segregation as explanations of the asymmetric population structure in the elderly phylogenies, the next hypothesis to test was whether clone-specific factors could lead to differential expansion among HSCs. Positively selected driver mutations were of key interest, due to the identification of known driver mutations as causal in a minority of the expanded clades. However, other possible causes of clone-specific expansion include epigenetic changes, telomere shortening and microenvironmental changes with age.

To look for evidence of pervasive positive selection in the dataset, an unbiased genetic analysis, known as dN/dS, was used[50]. The basis for this analysis is that synonymous mutations can be considered selectively neutral. Therefore, the rate of synonymous mutations (dS) can be compared to the rate of non-synonymous mutations (dN) within a dataset to identify if they occur at equivalent or different rates. A dN/dS ratio > 1.0 denotes an excess of non-synonymous mutations and provides evidence for positive selection, while a dN/dS ratio < 1.0 denotes a relative lack of non-synonymous mutations and provides evidence for negative selection. This approach can be carried out gene by gene (identifying specific genes under significant positive or negative selection) and across all coding genes (providing a global estimate of selection).

Novel driver genes

The gene-by-gene dN/dS analysis identified three genes under positive selection (**Fig. 4.6**). These were *DNMT3A* ($q=2.7\times10^{-11}$), *ZNF318* ($q=1.2\times10^{-6}$) and *HIST2H3D* (q=0.086).

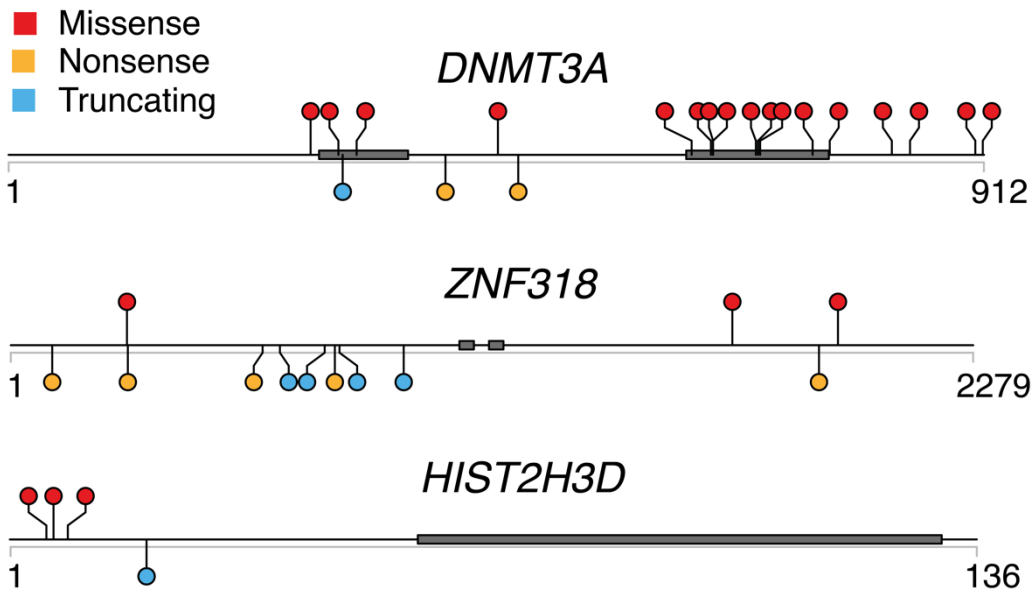


Fig. 4.6| Genes under positive selection in HSC/MPPs. Lolliplot plots to show the sites of variants in the dataset in the three genes under significant positive selection according to dN/dS. Thick grey bars denote locations of conserved protein domains.

DNMT3A is a well-known myeloid cancer gene that carried 23 mutations in our dataset, of which 13 were in expanded clades. In contrast *ZNF318* has not been well studied as a clonal haematopoiesis gene, but has been identified previously[82,177,178], and *HIST2H3D* has never previously been identified as under positive selection in blood (**Table 4.1**). ZNF318 is a zinc finger transcription factor, and carried predominantly truncating and nonsense mutations predicted to result in loss of function. Little is known about the function of the ZNF318 protein, but it has recently been shown to interact with TASOR, a pseudoPARP that recruits the HUSH complex to specific regions of DNA to promote DNA histone methylation (H3K9me3 via SETDB1)[179,180]. It therefore seems plausible that ZNF318 is a DNA reading 'adaptor' involved in epigenetic silencing of still unknown genes, and that its loss of function contributes to transcriptional changes promoting clonal expansion. *HIST2H3D,* a gene encoding a histone subunit, showed a cluster of missense mutations in the N terminal region, neighbouring amino acids whose post-translational modifications play a critical role in transcriptional regulation[181].

Table 4.1: Mutations identified in *HISTH3D* and *ZNF318*

PatientID	gene	Aachange	mutation_type
SX001	HIST2H3D	p.Q6K	missense
AX001	HIST2H3D	p.T7I	missense
AX001	HIST2H3D	p.R9C	missense
KX008	HIST2H3D	p.Q20*	truncating
AX001	ZNF318	p.W1959R	missense
KX003	ZNF318	p.S1709N	missense
KX003	ZNF318	p.S744*	truncating
KX003	ZNF318	p.Q768fs*22	nonsense
KX003	ZNF318	p.D278fs*23	nonsense
KX004	ZNF318	p.R931*	truncating
KX004	ZNF318	p.Q779*	truncating
KX004	ZNF318	p.S1914fs*8	nonsense
KX004	ZNF318	p.K597fs*3	nonsense
KX007	ZNF318	p.R638*	truncating
KX007	ZNF318	p.F99fs*41	nonsense
KX008	ZNF318	p.Y276C	missense

In order to investigate whether *ZNF318* and *HIST2H3D* mutations may play a role in myeloid malignancies, 534 AML genomes[145–147] were screened specifically for variants in *DNMT3A* and these two novel genes. While over 20% of AML cases carried mutations *in DNMT3A*, only one pathogenic *ZNF318* variant and no *HIST2H3D* variants were identified. This suggests that while variants in *ZNF318* and *HIST2H3D* are under selection in HSC/MPPs they do not necessarily contribute to the development of malignancy. *This analysis of AML genomes was performed by David Spencer*.

Global estimates of selection

The genome-wide estimate of dN/dS, calculated using the combined set of coding mutations from all individuals (25889 mutations), was significantly elevated at 1.06 ($CI_{95\%}$=1.03-1.09; **Fig. 4.7a**), providing evidence for significant positive selection in the dataset. The dN/dS approach can be influenced by biases in variant calling and nucleotide context, but the estimate was virtually unchanged after testing for potential biases (**Chapter 2**). We also performed a robust validation of the approach by generating a test dataset of mutation calls generated by random

insertion of coding variants (according to the spectrum of trinucleotide contexts observed in our data) into cord blood bam files that did not contain any coding mutations. The bam files then underwent exactly the same variant calling pipeline as the real data. This truly randomly generated test dataset reassuringly had a dN/dS ratio of 1.00.

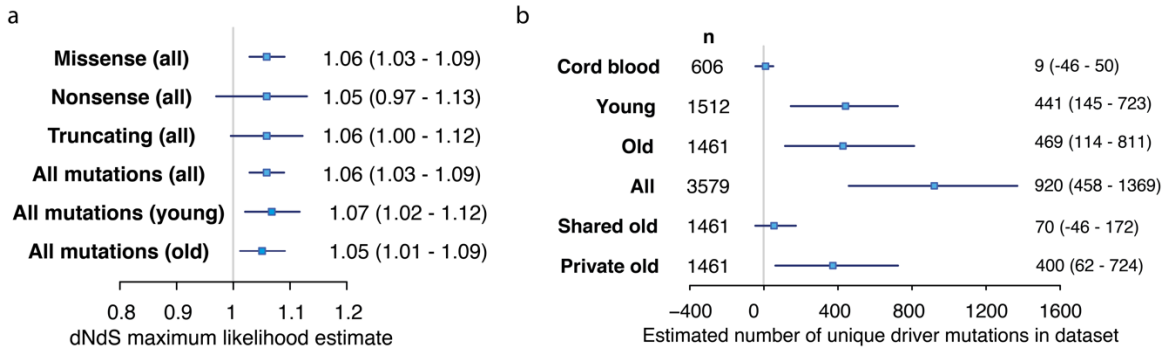


Fig. 4.7| dN/dS analysis. a, dN/dS maximum likelihood estimates for missense, nonsense, truncating and all mutations in the complete dataset (n = 25,888 coding mutations) and for all mutations in the young (individuals aged < 65 year) and old (individuals aged > 75 years) datasets analysed separately. The boxes show the estimate with whiskers showing the 95% CI. The numbers to the left give the numeric values for the estimates with 95%CI in brackets. **b,** Estimated number of driver mutations in the different datasets. The boxes show the estimate with whiskers showing the 95% CI. The numbers to the left give the numeric values for the estimates with 95%CI in brackets. 'n' is the number of cells included in each dataset.

A dN/dS ratio of 1.06 ($CI_{95\%}$=1.03-1.09) equates to approximately 1 in 18 (1/34 – 1/12) non-synonymous mutations in the data set being under positive selection. Estimated dN/dS ratios were almost identical in young and old individuals (1.06 and 1.05 respectively), providing evidence that the fraction of non-synonymous mutations (approximately 5%) under positive selection does not change with age.

An estimate of the total number of driver mutations in the dataset can be made using the following formula, where n_{NS} is the observed number of non-synonymous mutations, ω_{NS} is the estimated dN/dS ratio and, n_D is the number of drivers:

$$n_D = \frac{(\omega_{NS} - 1)}{\omega_{NS}} n_{NS}$$

Using the global dN/dS ratio given above and the number of non-synonymous mutation in our dataset (16,536) we can calculate the number of driver mutations as (1.06-1)*16536 = 936. This equates to > 100 drivers per adult individual (**Fig. 4.7b)**, and is considerably higher than the number of non-synonymous mutations we could identify in myeloid cancer genes.

Loss of Y analysis

Loss of the Y chromosome occurred frequently in the older males in our cohort. LOY has been previously identified in bulk blood samples from elderly men, and found to correlate with all-cause mortality[160]. The phylogeny approach used in this work, allowed the novel observation that LOY was occurring in multiple independent clones within an individual. Most strikingly the oldest male (KX003, 81 year male) had at least 62 independent LOY events across his phylogeny. A Monte Carlo test approach (**Chapter 2**) was used to show that LOY was significantly correlated with clonal expansion (p< 0.001; **Fig. 4.8**). Many of the expanded clades in our dataset had LOY, even in the absence of driver point mutations (**Figs. 3.8 and 3.9**). These LOY events could be timed to the first half of life. Together these data suggest LOY is under positive selection in HSCs, in keeping with the finding in mouse models that KDM6C on the Y chromosome suppresses leukaemogenesis[182] and recent work based on VAF spectrums of LOY in population cohorts[183].

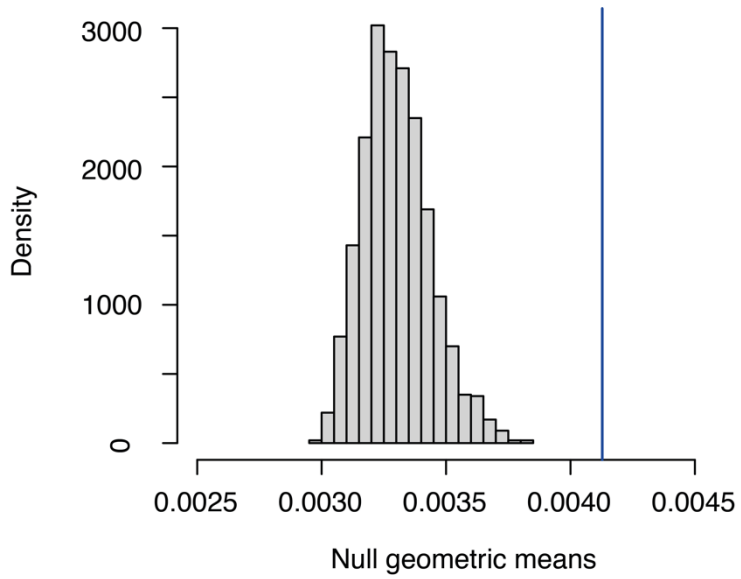


Fig. 4.8| Loss of Y is under positive selection in blood. Results of a randomisation / Monte Carlo test to define the null expected distribution of clade size for cells with loss of Y. This null distribution of geometric means from 2000 simulations is shown (histogram) together with the observed

geometric mean of clades with Y loss (vertical blue line). The observed value significantly outlies the expected distribution showing that clades with Y loss are significantly larger than would be expected by chance.

Summary evidence for positive selection

1. Gene specific dN/dS identified *DNMT3A*, *ZNF318* and *HISTH3D* to be under positive selection in the dataset.
2. The global estimate of dN/dS was 1.06 ($CI_{95\%}$=1.03-1.09), equating to 5% of non-synonymous mutations in the dataset being drivers (approximately 100 drivers per adult individual).
3. The phylogeny data provides evidence that LOY is under positive selection in HSCs.

4.5 Driver modelling

Driver modelling approach

The genetic analysis using dN/dS described above provides evidence for the existence of positive selection acting on previously unknown coding drivers in the dataset in both young and old individuals. The next step was to address if positive selection on unknown driver mutations could completely explain the observed HSC population structures in the elderly phylogenies. To do this, Approximate Bayesian Computation (ABC) modelling was used to infer HSC dynamics across the lifespan when 'driver mutations' entered the population according to varying parameters. Based on earlier results (**Chapter 3**) the HSC population size (*N*) was fixed at 100,000 and the time between symmetric cell divisions (τ) was fixed at 1 per year (after an initial population growth phase in the first few years of life).

The three parameters under investigation that were varied between simulations were:

1) Rate at which driver mutations enter the HSC population across life.
2) The 'rate' of the gamma distribution from which the fitness effects of the driver mutations was drawn.
3) The 'shape' of the gamma distribution from which the fitness effects of the driver mutations was drawn.

The ‘rate’ and ‘shape’ together define a distribution of fitness effects of driver mutations for a given simulation. The fitness effect (*s*) is defined as the average excess growth rate per year of a clone with a driver over that of wild-type HSCs (*s*= 0 indicates neutrality; **Fig. 4.9**). Initial simulations showed that the mutations with a fitness effect s>5% are unlikely to expand to >1% of HSCs over a human lifespan (**Fig. 4.9c**), so were excluded from the model.

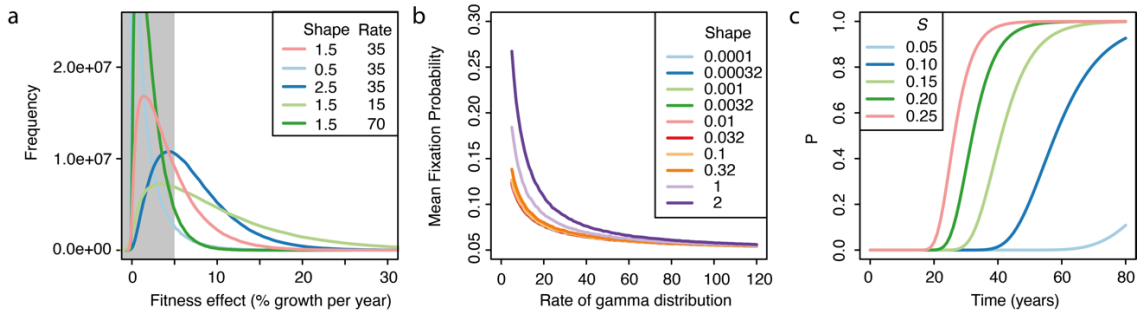


Fig. 4.9| Driver modelling parameter selection. a, Plot showing maximum posterior density estimates of the rate and shape parameters of the gamma distribution for selection coefficients (pink line) obtained using Approximate Bayesian computation. Blue/green lines show how altering the rate and shape parameters affect the gamma distribution. **b,** Plot showing how changing the shape of the gamma distribution of selection coefficients (each line has a different shape) alters the probability of a driver gene fixing in the population. Reducing the shape below 0.1 does not affect the probability of driver gene fixation and therefore was the lower limit of the shape prior. **c**, Plot showing how the probability of detecting a clone with CF 2.5% changes over time for different selection coefficients. There is only a probability of 0.1 of being able to identify a driver mutation with a selection coefficient of 0.05 that entered the population at birth. We therefore used a lower threshold of 0.05 for the driver mutation selection coefficients

The aims of the modelling approach were to firstly determine if any combination of the above parameters could generate simulations that were able to recapitulate the observed phylogenies at all ages. The second aim was to generate estimates for the parameters listed above, for which no previous estimates were available in the literature. **Figure 4.10** and **Chapter 2** provide more detail on the modelling approach used.

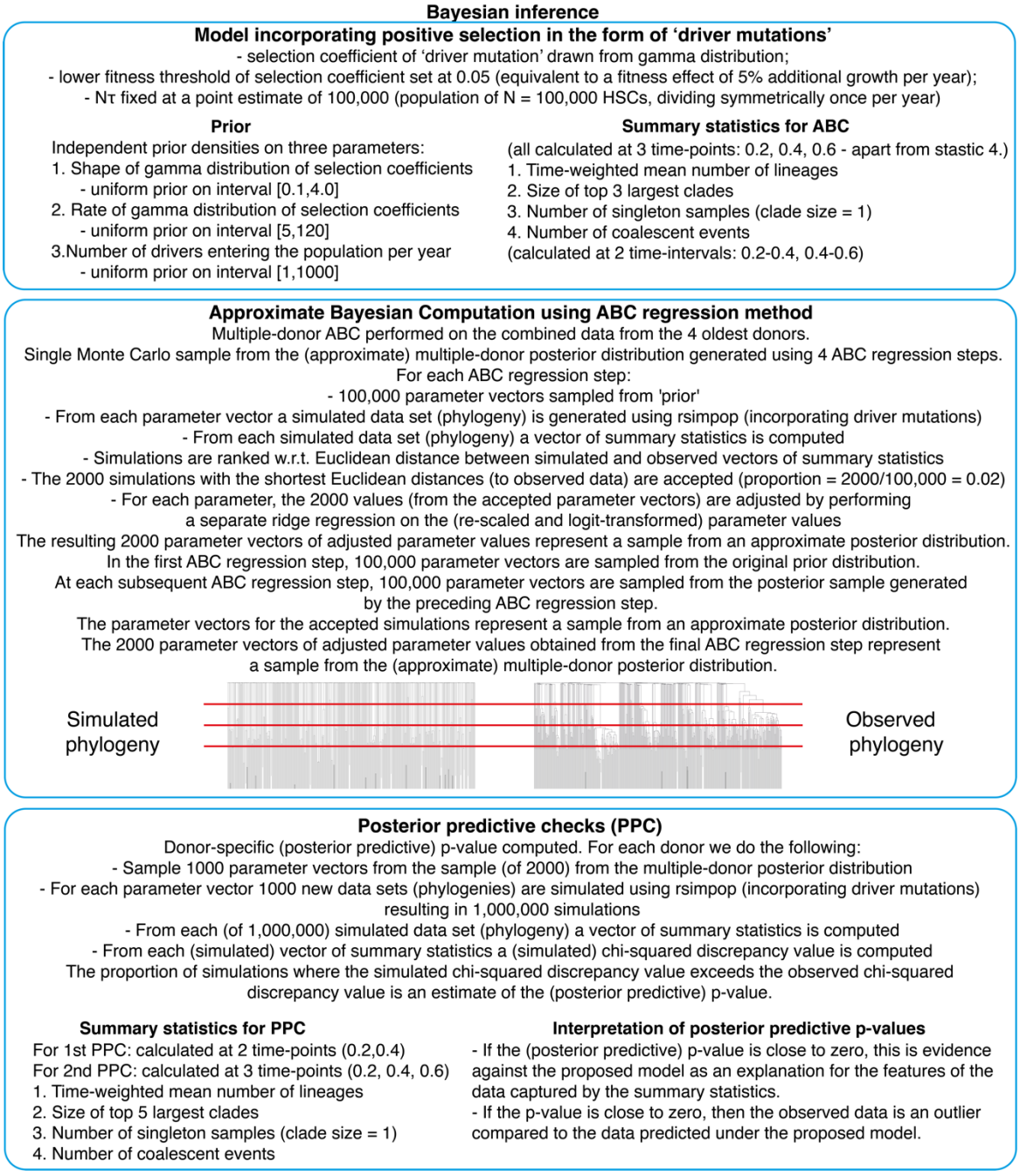


Fig. 4.10| Modelling of HSC populations incorporating positive selection. Overview of modelling approach used to estimate the shape and rate of the gamma distribution of selection coefficients from which 'driver mutations' are drawn, and the number of driver mutations drawn from this distribution (using a selection coefficient threshold of > 0.05) that are entering the HSC population per year. For these simulations $N\tau$ was fixed at 100,000 and therefore only summary statistics for the first 3 timepoints were used to assess how well a given simulation for an individual resembled the observed tree.

Driver modelling results

The ABC modelling described above showed that the observed trees could be accurately recapitulated at all ages (**Fig. 4.11**).

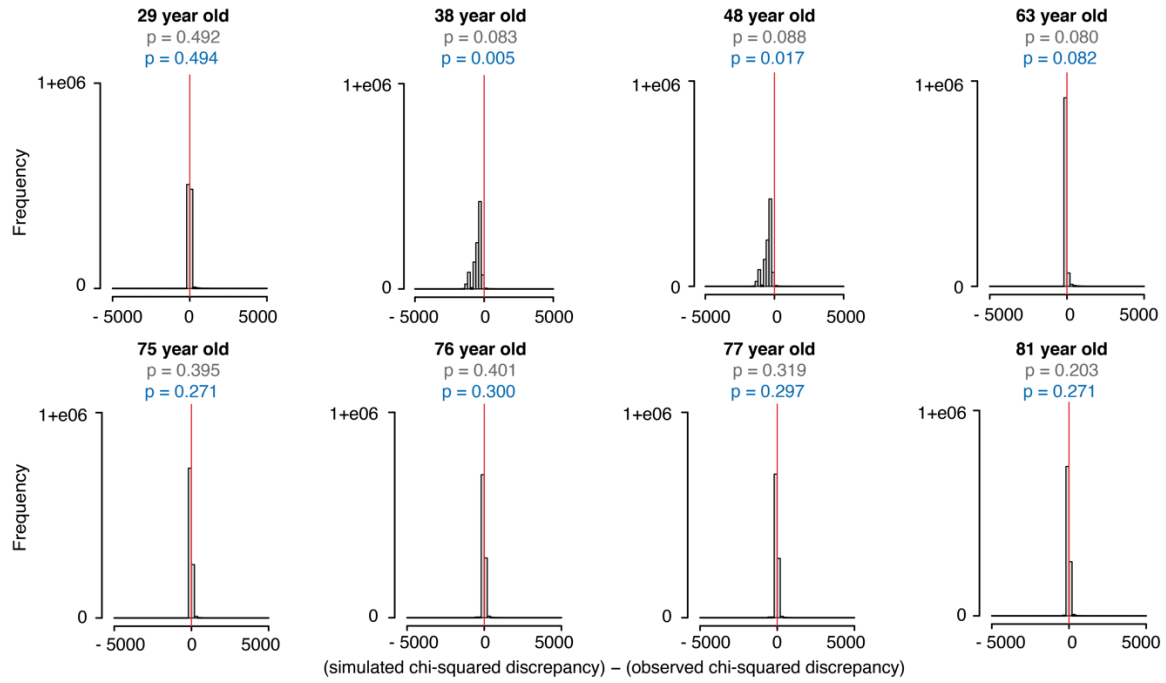


Fig. 4.11| Modelling of HSC populations incorporating positive selection. Plots showing the posterior predictive distribution of the difference between the predictive (simulated) chi-squared discrepancy and the observed chi-squared discrepancy, for each donor individual under the simple positive selection model. For the definition of the chi-squared discrepancy, and details of how the posterior predictive p-values are estimated, see Supplementary information "Posterior predictive model checking (PPC) methods which can be applied to Approximate Bayesian Computations (ABC)", Sections 1, 2 and 5. In these plots, the chi-squared discrepancy is computed from summary statistics evaluated at the first 3 (out of 4 equally spaced) timepoints on the phylogeny obtained from the specified donor (**Fig. 4.10**). For each donor, the posterior predictive distribution of the difference between predictive (simulated) and observed chi-squared discrepancy is represented as a histogram based on a Monte Carlo sample of at least 100,000 simulated phylogenies, drawn from the posterior predictive distribution. The proportion of simulated phylogenies which lie to the right of zero (red line) is a Monte Carlo estimate of the posterior predictive p-value (the probability that the predictive chi-squared discrepancy exceeds the observed chi-squared discrepancy under the positive selection model). Those p-values written in grey text are based on chi-squared discrepancies computed from summary statistics evaluated at the first 2 (out of 4 equally spaced) timepoints. Notice that these p-values are all above the 0.05 threshold, indicating that observed phylogenies (up to the second time point) are compatible with the simple positive selection model. Those p-values written in blue text are based on chi-squared discrepancies computed from summary statistics evaluated at the first 3 (out of 4 equally spaced) timepoints. Notice that all but two observed phylogenies (up to the third time point) are compatible with this positive selection

model. These p-values indicate that, once the third time point is included, the phylogenies of two of the younger individuals (38 year-old and 48 year-old) are no longer compatible with the positive selection model. Notice that these two donors also exhibit the most striking increase in population size from the middle part of the population trajectory onwards (**Fig. 3.10**). When all four timepoints are included, the phylogenies of 5 out of 8 donors have become incompatible with the positive selection model (data not shown). Only the phylogenies from the donors of ages 77, 76 and 29, remain compatible with the positive selection model. This suggests that the current positive selection model does not adequately account for the population processes towards the time of sampling.

The optimal simulated phylogenies had the same asymmetry of clade size distributions and timing of onset of markedly oligoclonal haematopoiesis (**Fig. 4.12**).

Fig. 4.12| Positive selection over life. Four consecutively simulated phylogenies of 380 cells sampled from a population of 100,000 cells that has been maintained at a constant $N\tau$ over life,

with incorporation of positively selected 'driver mutations'. The driver mutations have a fitness effect > 5% (drawn from a gamma distribution with shape = 0.47 and rate = 34) and enter the population at a rate of 200 per year. These are the maximum posterior density estimates of the rate and shape parameters obtained using the ABC method. The inclusion of these driver mutations is able to recapitulate a similar clade size distribution to that observed in the real HSPC phylogenies of the observed individuals across the whole age range. However, including driver mutations does not fully recapitulate the observed lack of coalescent events in the last 10-15 years of life, showing that an increase in $N\tau$ over this time is also required to fully recreate the patterns of coalescences in the real phylogenies. Driver mutations are marked with a symbol and their descendent clades are coloured. In all cases $N\tau$ is the same as the population size (N) as the generation time (τ) in all simulations is fixed at 1 year. The symbols / colours are not consistent for driver mutations between plots. The largest clades are therefore coloured in a consistent way beneath the plots to show how their size changes over time. The simulated phylogenies illustrate the complex clonal dynamics that can occur in later life as a result of clonal competition. While the majority of clades continue to expand, others stay relatively stable and some reduce in size. The phylogenies also show that by the age of 80 typically > 90% of HSCs in the population carry at least one driver mutation.

The elderly phylogenies contained sufficient information to provide meaningful refinement of the credible intervals for the acquisition rate and fitness effects of driver mutations (**Fig. 4.13**).

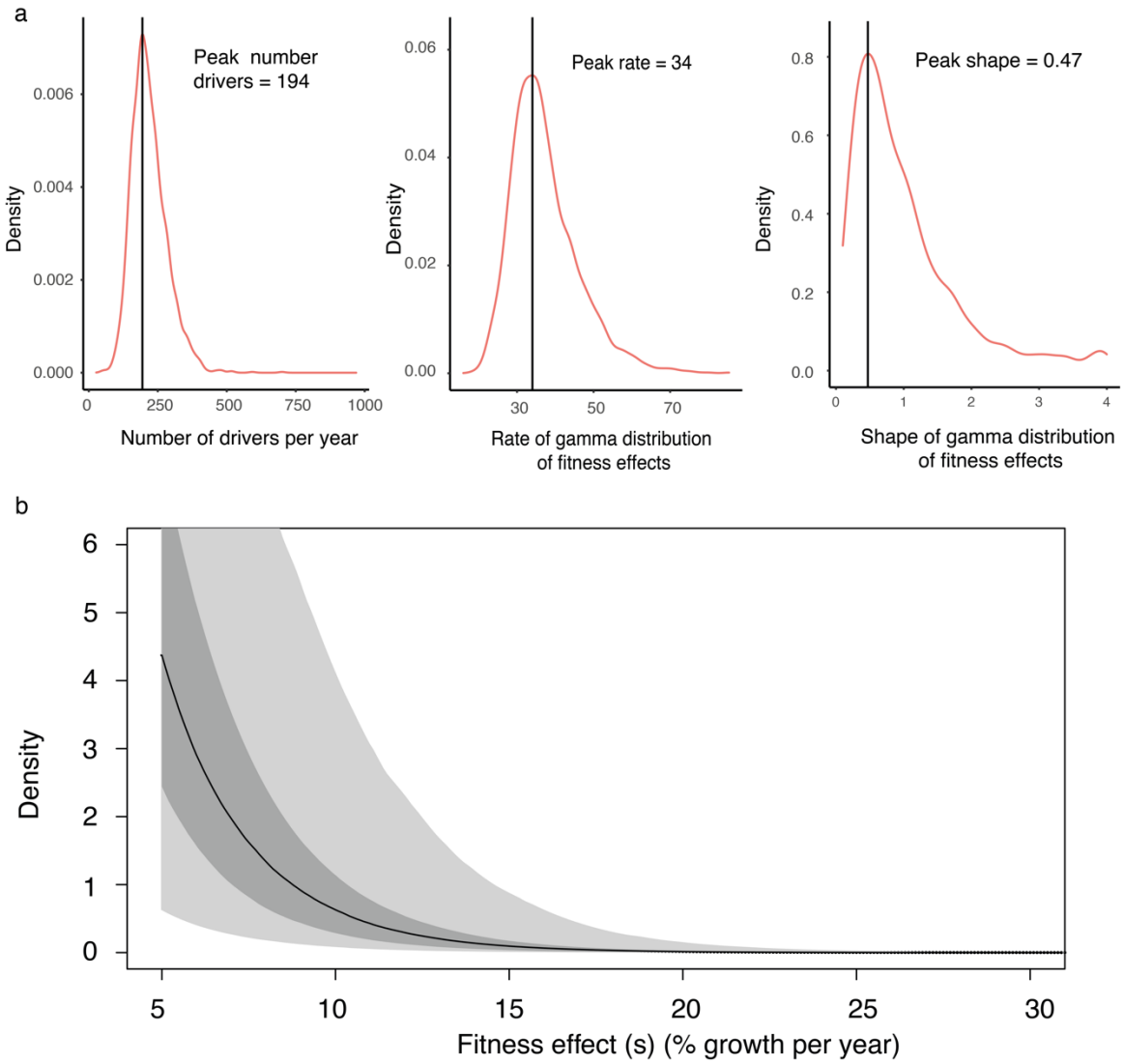


Fig. 4.13| Driver modelling parameter estimates. a, Posterior distributions for the three driver modelling parameters: 1) Number of 'driver' mutations with a fitness effect >5% entering HSC population of 100,000 cells per year, 2) Rate of gamma distribution of fitness effects, 3) Shape of gamma distribution of fitness effects. Black lines show peak estimates. **b**, Distribution of fitness effects for the 'driver' mutations entering the HSC population, as determined by ABC modelling approach. The point estimate for the shape and rate parameters of the gamma distribution were shape = 0.47 and rate = 34 (these are the univariate marginal maximum posterior density estimates shown in **a**). Interquartile range is shown in dark grey; 95% posterior intervals in light grey.

As shown above, driver mutations were estimated to enter the HSC compartment at a rate of 2.0×10^{-3}/HSC/year. The estimate obtained from the genetic analysis, based on dN/dS, was of drivers accruing at $3.6\text{-}10.0 \times 10^{-3}$/HSC/year (**Chapter 2**), which includes drivers with s<5% present in sequenced colonies. Therefore, even though the ABC estimates derive from branch

structures of the phylogenies and the genetic analysis relies on ratios of non-synonymous to synonymous mutations, the two approaches produce similar results. This implies much higher rates of driver mutation acquisition in blood than previously appreciated.

For driver mutation fitness effects, the gamma distribution most consistent with the data had a preponderance of moderate-effect drivers, with *s* in the range 5-10%, but a heavy tail of rare drivers conferring greater selective advantage (*s*>10%).

Implications of model

Having generated the optimal parameters for HSC driver mutation acquisition as described above, the model could be used to predict features of HSC populations that are impossible to accurately ascertain in the few phylogenies sampled. As a sanity check we could show that the Shannon Diversity Indices of 10,000 phylogenies generated using the optimal parameters from the modelling recapitulated the rapid drop off in clonal diversity between the ages of 60 and 70 as observed in the real phylogenies (**Fig. 4.14 compared to Fig. 4.1b**).

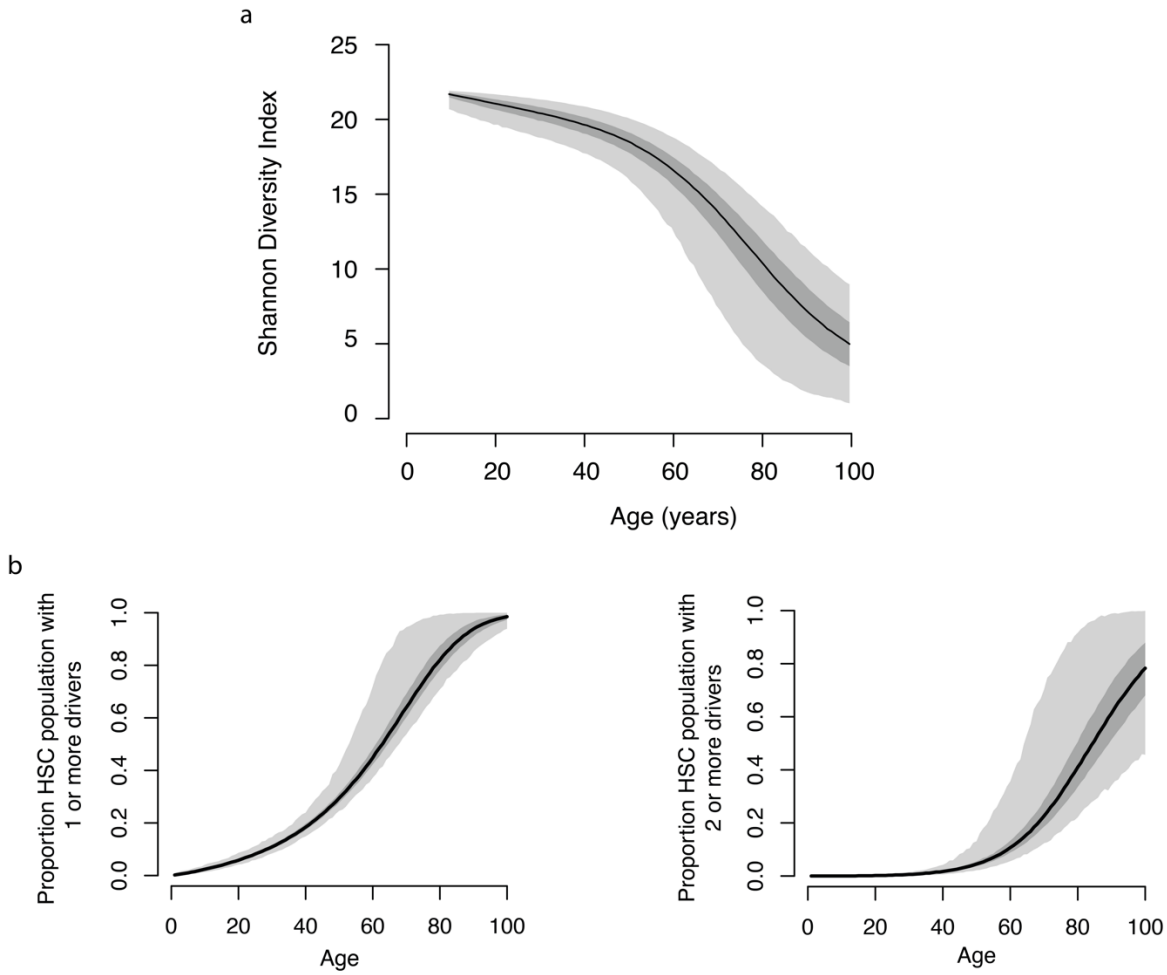


Fig. 4.14| Driver modelling Shannon diversity index and driver acquisition estimates. a, Plot showing the median, interquartile range, and 95% posterior intervals for Shannon Diversity Indices calculated yearly for 10,000 HSC population simulations run utilising the optimal parameter values for driver acquisition rate and fitness effects derived from the ABC modelling approach. The point estimate for the shape and rate parameters of the gamma distribution were shape = 0.47, rate = 34. The point estimate for the number of drivers with *s*>5% entering population per year = 200. **b,** Plot showing the median, interquartile range, and 1st and 99th percentiles for proportion of HSC population with drivers calculated yearly for 10,000 HSC population simulations run utilising the optimal parameter values for driver acquisition rate and fitness effects derived from the ABC modelling approach. The point estimate for the shape and rate parameters of the gamma distribution were shape = 0.47, rate = 34. The point estimate for the number of drivers with *s*>5% entering population per year = 200.

Using a similar approach, simulations could be used to predict that by the age of 80, 90% of HSCs harbour at least 1 driver mutation, and 50% of HSCs harbour 2 or more drivers (**Fig. 4.14b**). Given the universality, timing and extent of driver mutation acquisition in the HSC

population with age, it is possible that they could be contributing to ageing phenotypes in blood beyond the risk of cancer.

Summary driver modelling

1. Simulations incorporating driver mutation acquisition are able to accurately recapitulate HSC population structures at all ages.
2. Driver mutation acquisition rate (including only driver with *s*>5%) was estimated to be $2.0x10^{-3}$/HSC/year.
3. The distribution of driver fitness effects could be defined (many drivers with a fitness effect of 5-10% and a long tail of rare drivers with fitness effect >10%).
4. Simulations using the optimal parameter values suggest that by the age of 80, 90% of HSCs harbour at least 1 driver mutation and 50% harbour 2 or more.

4.6 Driver mutation timing and fitness effects

Driver mutation timing

The phylogeny structure and linear acquisition of somatic mutations through life, allows accurate timing of both known driver mutation acquisition, and the onset of expansion of clades with no known drivers (**Fig. 4.15**).

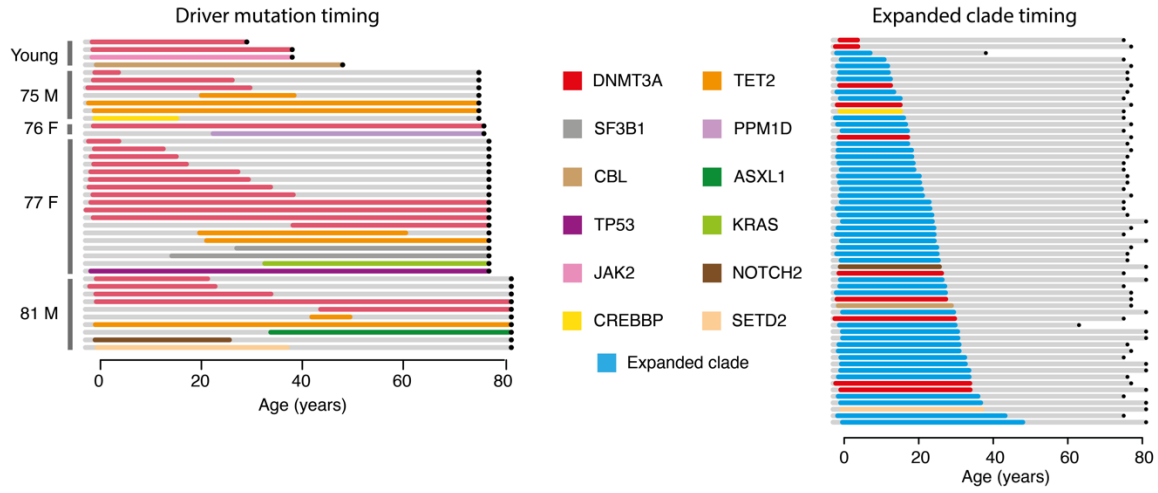


Fig. 4.15| Driver mutation timing. Plots illustrating the timing of acquisition of driver mutations and onset of clonal expansions respectively. Each line represents an expanded clade from an individual. The length of the line corresponds to the age of the individual and the coloured portion

represents the time of acquisition of the known or unknown driver events. These timings are inferred from the timing of the corresponding branches in the phylogenies depicted in **Figs. 3.8 and 3.9**. Bars are colours by gene mutation or blue for expanded clades with no known driver. Age 0 denotes the time of birth and black dots illustrate the age at sampling.

In both cases clonal expansions originate in either childhood or young adult life (in all cases before the age of 50). Of the 23 *DNMT3A* mutations in the dataset, 13 were in expanded clades. Of these, 2 were acquired before the age of 10, 3 before 20, and the remainder before 40 years. In almost all cases *DNMT3A* mutations time early in life, while many of the other known drivers in the dataset occur more commonly as 'second events' timing to later in life (for example *ASXL1, KRAS, PPM1D* and *SF3B1*). In contrast to phylogenies from individuals who developed myeloproliferative neoplasms[4], none of the driver or clonal expansion events could be timed to the '*in utero'* period.

Driver mutation fitness effects

For the 46 largest observed clones, fitness effects were directly estimated from the patterns of coalescences within their clade. This was performed using an algorithm *phylofit*, which can be thought of as a parametric adaptation of the *phylodyn* model (**Chapter 2**). *Phylofit* analysis produced estimates of *s* in the range of 10-30% (**Fig. 4.16)**, with expanded clade notation used shown in **Figure 4.17**. These results are consistent with the heavy tail of the credible gamma distribution of fitness effects from simulation models being in a similar range. The results also demonstrate that clones without known drivers can evolve comparable selective advantages to those with classic driver mutations.

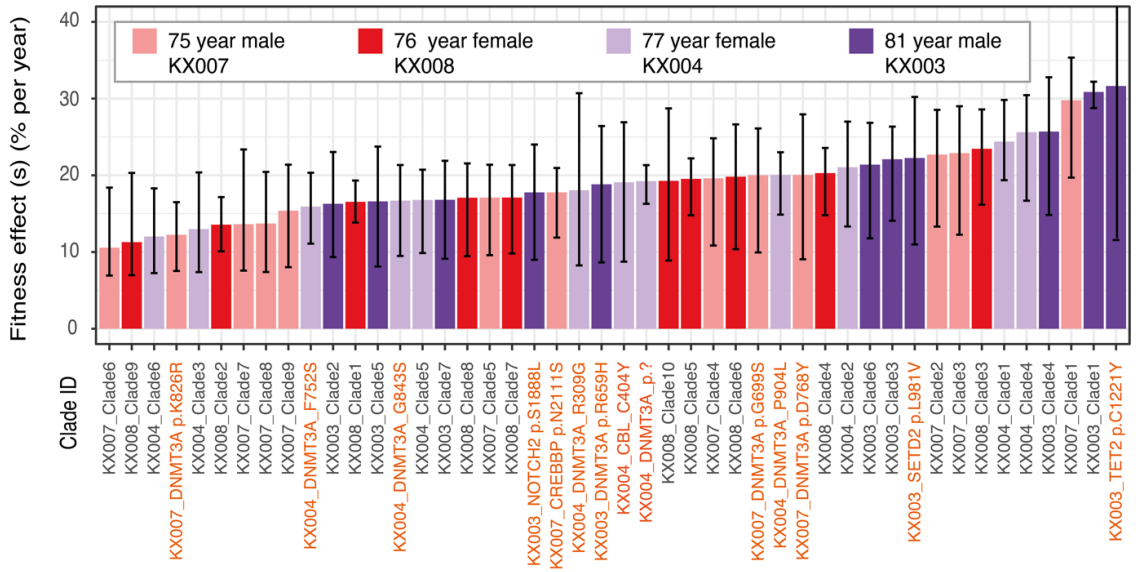


Fig. 4.16| Driver mutation fitness effects. Fitness effects within the HSC/MPP compartment are estimated for clades with causal driver mutations containing 4 or more HSC/MPP colonies (percent additional growth per year). Fitness effects are also estimated for expanded clades containing 5 or more HSC/MPP colonies (percent additional growth per year). Clade numbers are illustrated on the phylogenies in **Figure 4.17**.

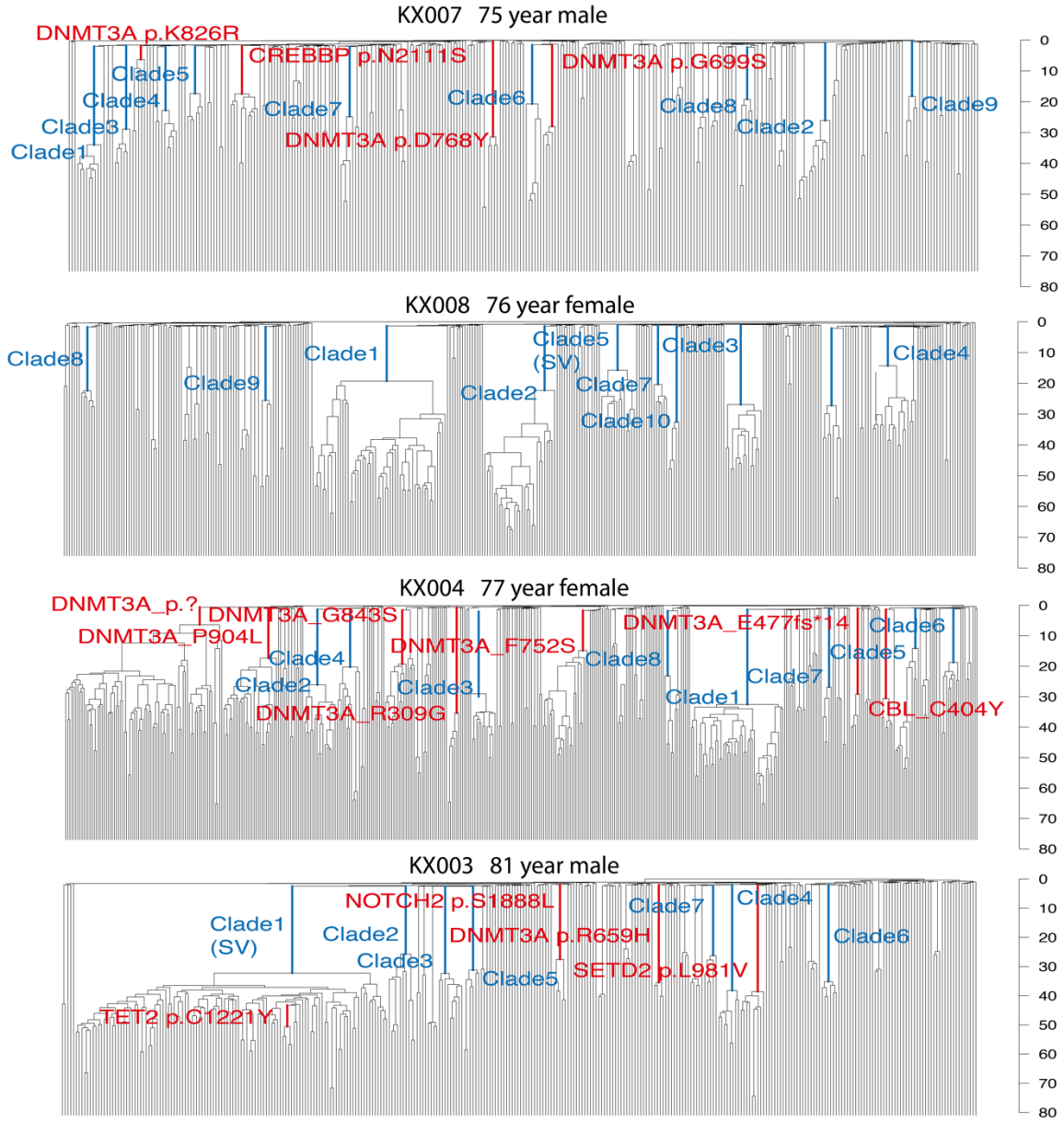


Fig. 4.17| Expanded clade annotations. Phylogenies of the four adults aged > 75 labelled with driver mutations and clade ID annotations as used in **Fig. 4.16**.

Putative drivers for clades with no known driver

To further investigate the hypothesis that coding mutations could be underlying the expanded clades with no known driver, a screen for putative coding drivers was undertaken. This was performed by looking for mutations in known cancer genes and the top 1% of dN/dS genes, as ranked by their gene-specific q-value. Using this approach, putative drivers could be identified for a significant proportion of the driverless expanded clades (**Fig. 4.18**). A caveat

to these results is that this approach is highly exploratory and that therefore these suggested drivers are only tentative predictions.

Clade	Fitness effect (%)	Possible drivers
KX001_Clade1	31 (11-54)	CDC26 p.R23*
KX002_Clade1	35 (13-51)	SPEN p.P2019H ACSM2B p.N193K
SX001_Clade1	NA	MAP3K1 p.T522fs*35
AX001_Clade1	34 (17-43)	KMT2D p.A4236E
KX007_Clade1	29 (19-35)	ST6GALNAC1 p.R590C
KX007_Clade3	22 (12-29)	FGFR1 p.M307I
KX007_Clade6	11 (7-18)	CSNK1A1 p.A216G
KX008_Clade3	23 (15-28)	PPP6C p.A189V LSP1 p.P61T

Clade	Fitness effect (%)	Possible drivers
KX008_Clade4	20 (14-23)	NOTCH3 p.D317N
KX008_Clade5	19 (14-22)	PPIP5K2 p.E940fs*12
KX008_Clade8	17 (9-21)	CIITA p.G516R
KX008_Clade9	11 (7-20)	RFWD3 p.P21fs*15
KX004_Clade1	24 (19-29)	KCNMA1 p.R909Q
KX004_Clade3	25 (17-30)	ZNF331 p.V343I
KX004_Clade5	12 (7-18)	ZNF318 p.Q779*
KX004_Clade6	16 (10-21)	KPNB1 p.S50fs*15
KX003_Clade6	21 (12-27)	LPHN2 p.E1290K

Fig. 4.18| Putative drivers. Table showing putative identified drivers for all four expanded clades in the young adult individuals (clade size 3 – 5). There are strong candidate drivers for these four individuals, whose expansions all have relatively high fitness effects, as would be expected for them to have been able to expand to a detectable level by a young age. The fitness effect was not calculated for SX001_Clade1 due to the clade size being 3 (too small for a meaningful *phylofit* analysis). For the elderly adult individuals, a clade size cut off of ≥ 5 was used as clade sizes of up to 4 can occasionally be seen under a neutral model by this age.

Driver mutation timing and fitness effects summary

1. In all cases clonally expanded driver mutations and the origins of clonal expansions with no known driver time to the first 5 decades of life, with the majority originating in childhood and young adult life.
2. Analysis of known and unknown driver mutation fitness effects reveals in all individuals the ‘s’ for their fittest clones is in the range 10-30% (in line with the results of the gamma distribution of fitness effects derived from driver modelling).
3. Putative driver mutations, in known cancer genes or the top 1% of genes identified in the gene-by-gene dN/dS analysis, could be identified for some driverless expanded clades.

4.7 Evidence for functional effect of known and unknown driver mutations

Colony size

To investigate the effect on replicative potential of the known and unknown driver mutations underlying clonal expansions, colony size was compared between the cells of different clades within each elderly individual (**Fig. 4.19**). To maximise the power for this analysis immunophenotyping data from which colony size could be calculated was split into a maximum of four groups per individual: 1) Largest clade; 2) Other *DNMT3A* clades; 3) Other expanded clades and 4) Non-expanded (all other samples). Only data from HSC-derived colonies with more than 200 myeloid cells immunophenotyped were included in the analysis (see **Chapter 2** for the gating strategy used for colony immunophenotyping). All colonies underwent immunophenotyping at 3 weeks, other than a subset of colonies from KX003 (who had the lowest colony efficiency and smallest colonies), for whom phenotyping was also undertaken at 4 weeks on colonies that had grown large enough in size by then.

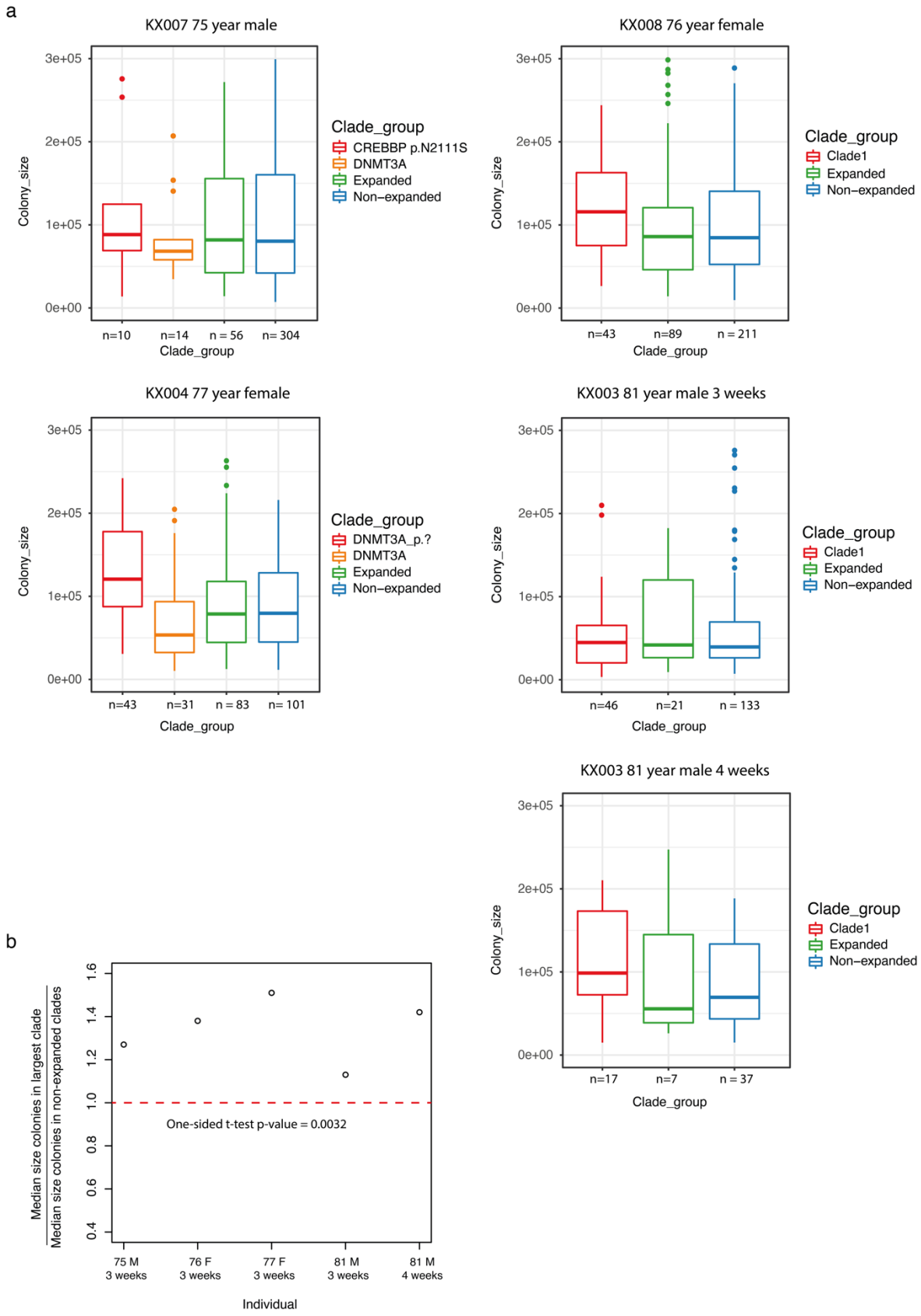


Fig. 4.19| Colony size by clade type. a, Boxplots for each adult aged >70 showing total colony size by clade in the following groups: 1) Largest clade; 2) *DNMT3A* expanded clades (if applicable,

excluding largest clade); 3) Other non-*DNMT3A* expanded clades; 4) Non-expanded clades (all other samples). All colony sizes assessed at 3 weeks other than for KX003 where colony size for colonies greater than approx. 3000 cells in size was assessed at both 3 weeks (and for smaller colonies that continued to grow in size, at 4 weeks). **b**, Median size myeloid colonies in largest clade divided by median size myeloid colonies in non-expanded clades plotted for each elderly adult individual. One-sided t-test confirms that colonies from the largest clade in each individual are significantly larger than those from non-expanded clades. Red dotted line shows equality of size between largest clade and non-expanded clades (1 = no difference in median size).

As shown in **Figure 4.19a**, there was a trend for colonies from the largest clade in each individual to be larger than the colonies from 'non-expanded' clades. This trend met statistical significance using a Wilcoxon Rank Sum Exact Test for three individuals: KX008 (76 year female, p-value = 0.030), KX004 (77 year female, p-value = 0.001) and the 4 week timepoint for KX003 (81 year male, p-value = 0.042). There was also a trend for 'DNMT3A' mutated colonies to be smaller than colonies from 'non-expanded' clades but this did not meet statistical significance. **Figure 4.19b** compares the median size of colonies from the largest clade in each individual with the median size of colonies from their non-expanded clades. A one-sided t-test was used to confirm that overall, colonies from the largest clades are larger than those from non-expanded clades (p-value = 0.0032). This analysis of colony size suggests that known and unknown driver mutations underlying the largest clonal expansions in elderly individuals can affect replicative potential.

Colony phenotype

To investigate the effect on lineage bias of the known and unknown driver mutations causing clonal expansions, colony phenotypes were compared between the four groups per individual described above: 1) Largest clade; 2) Other *DNMT3A* clades; 3) Other expanded clades and 4) Non-expanded (all other samples) (**Fig. 4.20**). Exactly the same criteria as for the colony size assessment were used for inclusion in the colony phenotype by clade analysis.

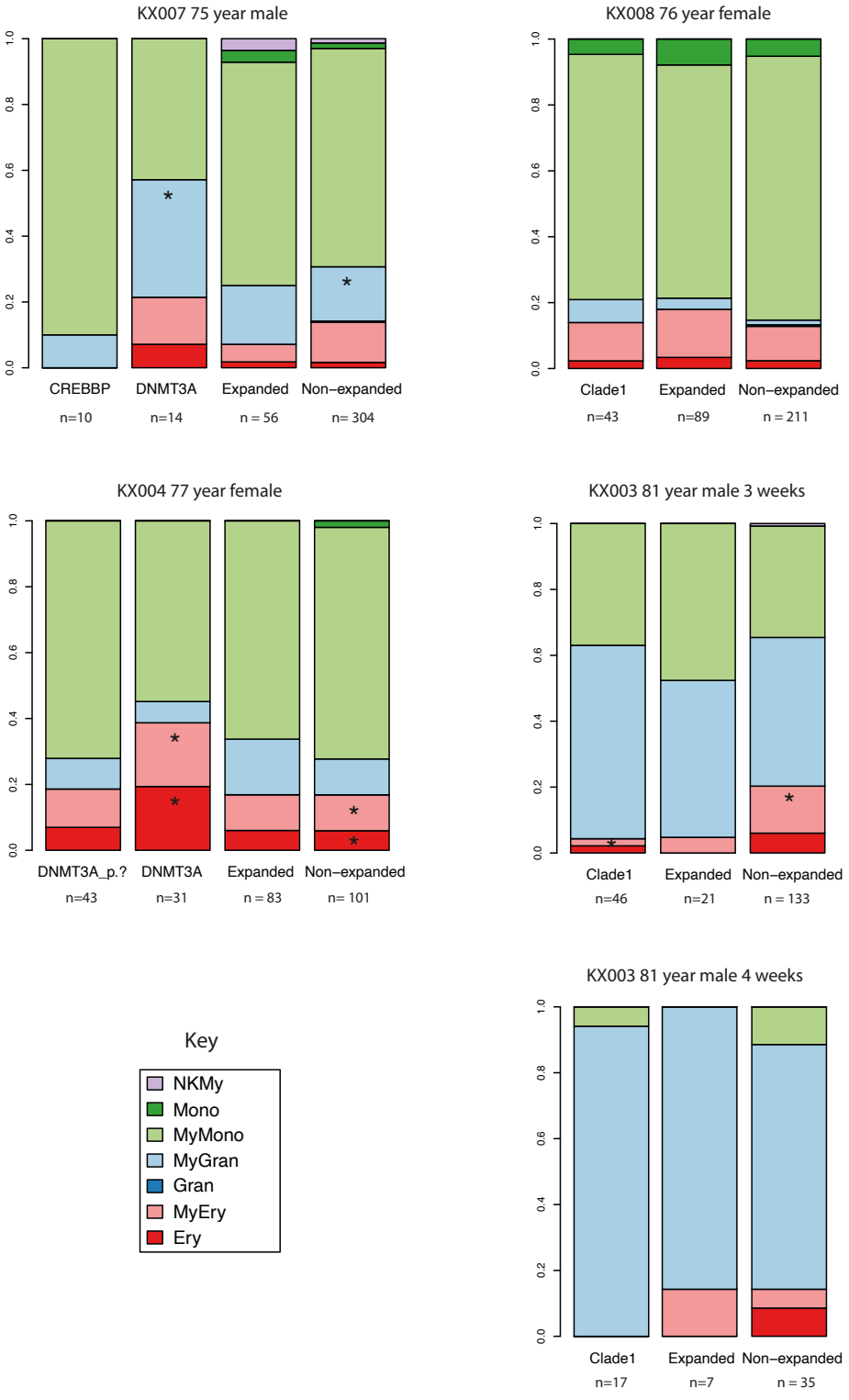


Fig. 4.20| Colony phenotype by clade type. Barplots for the four adults aged >70 years showing colony phenotype by clade in the following groups: 1) Largest clade; 2) *DNMT3A* expanded clades (if applicable, excluding largest clade); 3) Other non-*DNMT3A* expanded clades; 4) Non-expanded

clades (all other samples). All colony phenotypes for colonies >3000 cells in size assessed at 3 weeks other than for KX003 where colony size was assessed at both 3 weeks (and for smaller colonies that continued to grow in size, at 4 weeks). Ery = erythroid cells only; EryMy = erythroid cells and myeloid cells (any combination of monocytes and granulocytes); Gran = granuloctes only; MyGran = granulocytes and monocytes (granulocytes predominant); Mono = monocytes only; MyMono = monocytes and granulocytes (monocytes predominant); NKMy = NK cell and myeloid cells (any combination of monocytes and granulocytes); NKMy = NK cell and myeloid cells (any combination of monocytes and granulocytes). Fisher test used to determine significant difference in colony phenotype. Star (*) denotes significant difference (p-value < 0.05) for summed phenotypes below and including the starred phenotype when compared to non-expanded clades.

The first main conclusion from the colony phenotyping analysis is that *DNMT3A* mutations typically result in increased erythroid cell output, with both Ery and EryMy phenotypes present at increased fraction in the *DNMT3A* mutant colonies of KX007 and KX004 (KX004 Fisher test p-value = 0.01). This finding is in keeping with recent studies that have shown expansion of *DNMT3A*-mutated immature myeloid progenitors primed toward megakaryocytic-erythroid fate in both mouse models[184] and individuals with *DNMT3A* R882 clonal haematopoiesis[185]. However the largest *DNMT3A* mutated clade in KX004 (*DNMT3A* p.?, an essential splice site mutation), did not confer an increased erythroid cell output. This could be in keeping with a different phenotype of the splice site mutation or the presence of additional 'unknown' drivers that may modify the mutant *DNMT3A* effect. Given that this is the largest clade in KX004, the presence of additional drivers is highly likely based on the driver modelling results discussed earlier.

The second main conclusion is that in the case of two of the other largest clades (the *CREBBP* mutated clade in KX007 and Clade1 in KX003) there is a reduction in erythroid cell output, meeting significance for Clade1 in KX003 (Fisher test p-value = 0.01). While preliminary, these phenotypic findings confirm that the largest expanded clades within an individual commonly possess altered function in terms of lineage bias and mature cell output.

Summary functional effects of driver mutations

1. Colonies from the largest clade were significantly larger than those of 'non-expanded' lineages in three out of four elderly individuals (although for KX003 only at the 4-week immunophenotyping timepoint).

2. There was a trend for *DNMT3A* mutated colonies to be smaller than colonies from 'non-expanded' lineages, although this did not meet statistical significance.
3. *DNMT3A* mutations typically result in increased erythroid output, as found in KX004 and KX007.
4. Two of the other largest clades (in KX007 and KX003) show an opposite pattern, with a reduction in erythroid containing phenotypes amongst the immunophenotyped colonies.

4.8 Summary

In this chapter, the mechanisms underlying the profound change in clonal structure of HSC phylogenies in the elderly have been explored. The data presented provide considerable evidence in support of positive selection on many previously unknown drivers being the most likely explanation. Results from the genetic analysis using dN/dS and ABC modelling of driver mutation acquisition are remarkably consistent in terms of estimates for the number of driver mutations entering the HSC population (approximately 2.0×10^{-3}/HSC/year with $s > 5\%$). By the age of 80, the driving modelling would predict pervasive acquisition of driver mutations, with approximately 90% of HSCs harbouring at least 1 driver mutation. The immunophenotyping results provide evidence that the known and unknown driver mutations in the largest expanded clades do alter mature cell output in terms of both replicative potential and lineage bias.

In conclusion, pervasive driver mutation acquisition and the resulting re-wiring of HSC cellular pathways, likely contributes to loss of function in the haematopoietic system with age. The extent to which the findings presented in **Chapters 3 and 4** are altered in individuals exposed to chemotherapy are the focus of **Chapter 5**.

Chapter 5: Impact of chemotherapy on the haematopoietic system

5.1 Introduction

Little is known about the effect of chemotherapy on normal human HSCs, in terms of both the mutagenic insult and impact on HSC population size or clonal structure. There are several lines of evidence that suggest chemotherapy can have a significant impact on the normal HSC population in humans. First, there is an increased risk of haematological malignancy in individuals treated with chemotherapy[109,186,187]. Second, chemotherapy increases the overall incidence of clonal haematopoiesis and favours mutations in the DNA damage response genes *PPM1D*, *TP53* and *CHEK2*[106–108]. Third, chemotherapy commonly results in acute and prolonged cytopenias[188–191]. Fourth is the clinical observation that following intensive chemotherapy the haematopoietic system is often less resilient in the face of subsequent chemotherapeutic insult, particularly in older individuals. Fifth, murine studies have shown that the agent 5FU reliably recruits HSCs to proliferate and differentiate, showing that chemotherapy does affect HSC fate decisions in the short term[103,163]. However, the cellular mechanisms by which chemotherapy exerts these effects remain largely unclear, as does any distinction between the specific cellular or population level effects of different agents. A better understanding of the impacts of commonly used chemotherapeutic agents could allow us to reduce the toxicity of regimens, especially if interchangeable agents have very different mutagenic impacts on normal HSCs.

Key questions to be addressed in this chapter

1. What is the mutagenic impact of a range of chemotherapeutic agents on normal HSCs?
2. Is the mutagenic impact of chemotherapy uniform across chemotherapeutic agents in the same class?
3. Is the mutagenic impact of particular chemotherapeutic agents uniform across individuals?
4. Is the mutagenic impact of particular chemotherapeutic agents uniform across all cells sampled from a single individual?
5. Which mutational signatures are responsible for any increase in mutation burden due to chemotherapy in normal HSCs?

6. Does chemotherapy impact HSC population size?
7. Does chemotherapy impact HSC population clonal structure?

These questions were investigated by whole genome sequencing of peripheral blood single-cell derived HSPC colonies (grown in Methocult, **Chapter 2**) from individuals previously exposed to chemotherapy. The HSPC colony growth was performed by Anna Clay (research assistant) under my supervision. This work was performed as part of the 'Mutographs Project', a CRUK funded Grand Challenge Project and analysis of the data is still ongoing.

5.2 Clinical information and samples

Nineteen chemotherapy exposed patients were recruited from Addenbrooke's Hospital for this study (**Table 5.1 and 5.2**). The majority of the patients were treated for either lymphoma (recruited by myself and Dr Daniel Hodson), or bowel cancer, one of whom also had a diagnosis of myeloma (recruited by Dr Ultan McDermott). To investigate a wider range of platinum agents, we also included one lung cancer patient (recruited by Dr Gary Doherty) and one paediatric neuroblastoma patient (recruited by Dr Aditi Vedi). Eleven normal individuals were included in the control cohort (**Table 5.3**).

For the majority of chemotherapy exposed cases only 4-10 HSPCs were sequenced at high coverage (total 151 samples, mean sequencing depth 23X). For five individuals (1 with follicular lymphoma, 2 with Hodgkin's disease, 1 with bowel cancer and 1 with lung cancer) a larger number (41-259 HSPCs per individual) were sequenced. The samples for these five individuals were sequenced at lower coverage (total 558 samples, mean sequencing depth 15X). These larger phylogenies were generated to 1) investigate whether the mutagenic insults have a uniform impact across cells within an individual; 2) to determine if chemotherapy has an impact on HSC population structure.

The 'chemotherapy exposed' HSPC data was compared to data from a total of 11 normal individuals. The normal comparator dataset included a subset of data from 9 of the normal individuals studied in **Chapters 3 and 4** (all the adult donors and 1 cord blood donor). Two additional normal donors aged 60 and 63 were also included, to provide a better age matched

range to the ‘chemotherapy-exposed’ population. To avoid bias and to best match sequencing depth with that of the ‘chemotherapy exposed’ cohort, normal samples from each of these eleven individuals were ordered by sequencing depth and the 10 samples with the highest sequencing depth were selected for inclusion in the mutation burden part of the analysis (total 110 samples, mean sequencing depth 24X). For the HSC population structure analysis, age matched normal phylogenies of the same size were constructed for comparison with the larger phylogenies obtained from five individuals in the chemotherapy exposed cohort.

Table 5.1: Clinical information chemotherapy exposed cohort

PDID	Age / Sex	Cancer type[1]	Chemotherapy (years post chemotherapy)	Smoking (pack years)	No. samples mutation burden analysis[4]	No. samples population structure analysis
PD44579	63 F	FL	R-CVP (2)	No (0)	10	173
PD47539	64 F	FL	R-CVP (6)	No (0)	5	NA
PD47540	57 F	FL	R-CHOP (4)	No (0)	5	NA
PD47695	72 F	DLBCL	R-CHOP (5)	No (0)	5	NA
PD47696	74 M	DLBCL	R-CHOP (0)	No (0)	5	NA
PD46541	64 M	FL	R-Benda (0)	No (0)	10	NA
PD47703	48 F	HD HD	ChlVPP (37) BGeV (0)	No (0)	10	218[2]
PD50308	27 F	HD	EscBEACOPP (2)	No (0)	10	42
PD47699	79 F	M CC	VMP (4) + CTDa (2) Ox Cap (0)	No (0)	10	NA
PD47538	61 F	CC	5FU Irino (0)	Ex (12)	10	NA
PD47702	63 F	CC	5FU Ox (0)	No (0)	10	NA
PD47697	72 F	CC	Cap Ox (0)	Ex (8)	10	NA
PD47698	68 M	CC	Cap Ox (0)	Ex (20)	10	NA
PD47536	67 F	CC	Cap Ox 5FU Irino[3] (0)	No (0)	5	NA
PD47537	61 F	CC	Cap Ox 5FU Irino[3] (0)	Ex (35)	5	44
PD47700	80 M	CC	Cap Ox 5FU Irino[3] (0)	No (0)	4	NA
PD47701	68 F	CC	Cap Ox 5FU Irino[3] (0)	No (0)	10	NA
PD50307	40 F	LC	Carbo Etop (0)	Yes (50)	10	41
PD50306	3 M	NB	Rapid COJEC (0)	No (0)	9	NA

[1] FL = follicular lymphoma, DLBCL = diffuse large B cell lymphoma, HD = Hodgkin's disease, CC = colon cancer, M = myeloma, LC = lung cancer, NB = neuroblastoma

[2] An additional 41 samples from this patient were sequenced a year later after treatment for DLBCL with R-CHOP.

[3] These individuals were treated for metastatic CC with alternating regimens including these drugs over a number of years prior to sampling and therefore had higher exposures.

[4] All sequenced samples derived from PB HSPC colonies grown in methocult.

Table 5.2: Chemotherapy agent information chemotherapy exposed cohort

PDID	Age / Sex	Alkylating agent[1]	Platinum agent	Anti-metabolite	Topo-inhibitor	Vinca-alkaloid
PD44579	63 F	Cyclophosphamide	None	None	None	Vincristine
PD47539	64 F	Cyclophosphamide	None	None	None	Vincristine
PD47540	57 F	Cyclophosphamide	None	None	Doxorubicin	Vincristine
PD47695	72 F	Cyclophosphamide	None	None	Doxorubicin	Vincristine
PD47696	74 M	Cyclophosphamide	None	None	Doxorubicin	Vincristine
PD46541	64 M	Bendamustine	None	None	None	None
PD47703	48 F	Procarbazine Chlorambucil Bendamustine	None	Gemcitabine	None	Vinblastine Vinorelbine
PD50308[2]	27 F	Procarbazine Cyclophosphamide	None	None	Etoposide Doxorubicin	Vincristine
PD47699	79 F	Melphalan Cyclophosphamide	Oxaliplatin	Capecitabine	None	None
PD47538	61 F	None	None	5FU	Irinotecan	None
PD47702	63 F	None	Oxaliplatin	5FU	None	None
PD47697	72 F	None	Oxaliplatin	Capecitabine	None	None
PD47698	68 M	None	Oxaliplatin	Capecitabine	None	None
PD47536	67 F	None	Oxaliplatin	5FU / Cap[3]	Irinotecan	None
PD47537	61 F	None	Oxaliplatin	5FU / Cap[3]	Irinotecan	None
PD47700	80 M	None	Oxaliplatin	5FU / Cap[3]	Irinotecan	None
PD47701	68 F	None	Oxaliplatin	5FU / Cap[3]	Irinotecan	None
PD50307	40 F	None	Carboplatin	None	Etoposide	None
PD50306	3 M	Cyclophosphamide	Carboplatin Cisplatin	None	Etoposide	Vincristine

[1] All agents in this category are bifunctional alkylating agents (have two DNA binding reactive sites per molecule), other than procarbazine which is a monofunctional alkylating agent (single reactive site per molecule). This has implications for the ability of the drugs to crosslink DNA strands (only possible for bifunctional agents).

[2] This individual also received bleomycin, a cytotoxic antibiotic.

[3] Cap = capecitabine, 5FU = 5-Fluorouracil. Capecitabine is a prodrug of 5-Fluorouracil.

Table 5.3: Clinical information normal cohort

PDID	Age / Sex	Sample source	Clinical info	Smoking (pack years)	No. samples mutation burden analysis	No. samples population structure analysis
PD40315	0 F	CB HSC/MPP	Normal	No (0)	10	NA
PD40521	29 M	PB HSC/MPP	Normal	No (0)	10	42
PD40667	38 M	BM HSC/MPP	Crohn's disease	Ex (8)	10	41
PD41048	47 M	PB HSC/MPP	Selenoprotein deficiency	No (0)	10	218
PD49237	60 F	BM HSC/MPP	Normal	No (0)	10	44
PD42976	62 M	PB HSC/MPP	Normal	No (0)	10	NA
PD49236	63 M	BM HSC/MPP	Normal	No (0)	10	173
PD47738	75 M	BM HSC/MPP	Normal	Yes (8)	10	NA
PD48402	76 F	BM HSC/MPP	Normal	No (0)	10	NA
PD45534	77 F	BM HSC/MPP	Normal	No (0)	10	NA
PD43974	81 M	BM HSC/MPP	Normal	No (0)	10	NA

5.2 Mutation accumulation due to chemotherapy

Validation of approach

An adapted version of the variant filtering approach used for the ageing cohort samples in **Chapters 3 and 4**, was used for the chemotherapy exposed and normal 'mutation burden dataset' samples in this chapter (method described in detail in **Chapter 2**). The adapted approach was required due to the small number of samples per individual (≤ 10), which precluded use of binomial-based filtering. The normal data presented in the mutation burden analysis section was reanalysed using identical pipelines to the chemotherapy-exposed data. The adapted filtering approach was validated by comparing results from normal samples filtered both using the adapted filtering strategy, which only included 10 samples per individual, and the binomial filtering strategy which included all samples from that individual in the larger ageing dataset (**Fig. 5.1a)**. Reassuringly these results were highly concordant. Whether differences in mutation burden could be due to the fact that the normal samples were HSC/MPP derived, while the chemotherapy exposed samples were HSPC derived was also considered. Mutation burden analysis of normal HSPCs presented in **Chapter 3** identified no difference in mutation burden between HSC/MPP derived and HPC derived colonies grown in liquid culture (**Fig. 3.2b**). Comparison of normal HSC/MPP liquid culture colony mutation

burdens with Methocult HSPC derived samples from four normal individuals also found no difference in mutation burden (**Fig. 5.1b**). Mean sequencing depth was comparable across the normal and chemotherapy cohorts (**Fig. 5.1c**), the normal cohort having a slightly higher mean sequencing depth (24X) than the chemotherapy exposed cohort (23X).

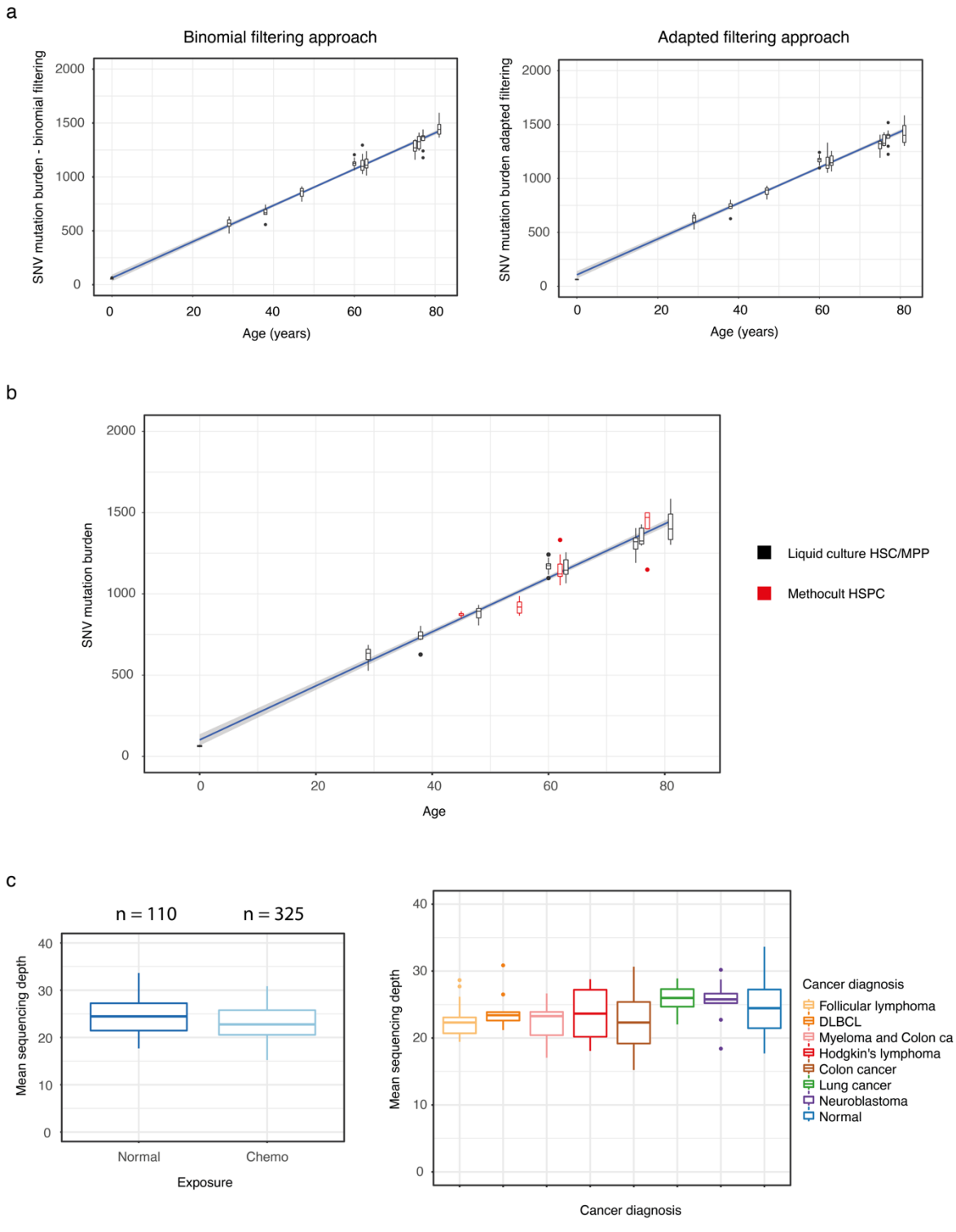


Fig. 5.1| Validation of variant filtering approach. a, SNV mutation burden of 10 samples per normal individual with variants filtered using the binomial filtering strategy that made use of several hundred other samples sequenced from the individual (left). SNV mutation burden of the same 10 samples per normal individual with variants filtered using the adapted filtering strategy that required only 10 samples per individual (right). **b**, Comparison of SNV mutation burden between normal HSC/MPP colonies grown in liquid culture (black) and normal HSPC colonies grown in Methocult (red). **c**, Mean sequencing depth per sample in the normal vs chemotherapy-exposed cohorts (right) and by diagnosis (left).

Single nucleotide variant burden in chemotherapy exposed individuals

The SNV mutation burden varied enormously across the chemotherapy exposed cohort, with some HSPCs harbouring thousands of excess mutations compared to other HSPCs with a completely normal mutation burden (**Fig. 5.2a,b**). There were three individuals whose HSPCs had accumulated over a thousand excess SNVs compared to that expected for an age-matched normal individual. One of these was a paediatric patient aged 3 treated for neuroblastoma, who had a mutation burden greater than that of a normal 80 year old. The other two were both individuals treated for Hodgkin's lymphoma. The HSPCs from these three individuals also showed a dramatically elevated variance in mutation burden as compared to the normal individuals (with a range in mutation burden of over 1500 mutations in one individual). Furthermore, in these high mutation burden cases, none of the cells showed a normal mutation burden, suggesting all normal HSPCs in these individuals were affected by chemotherapy to a greater or lesser extent.

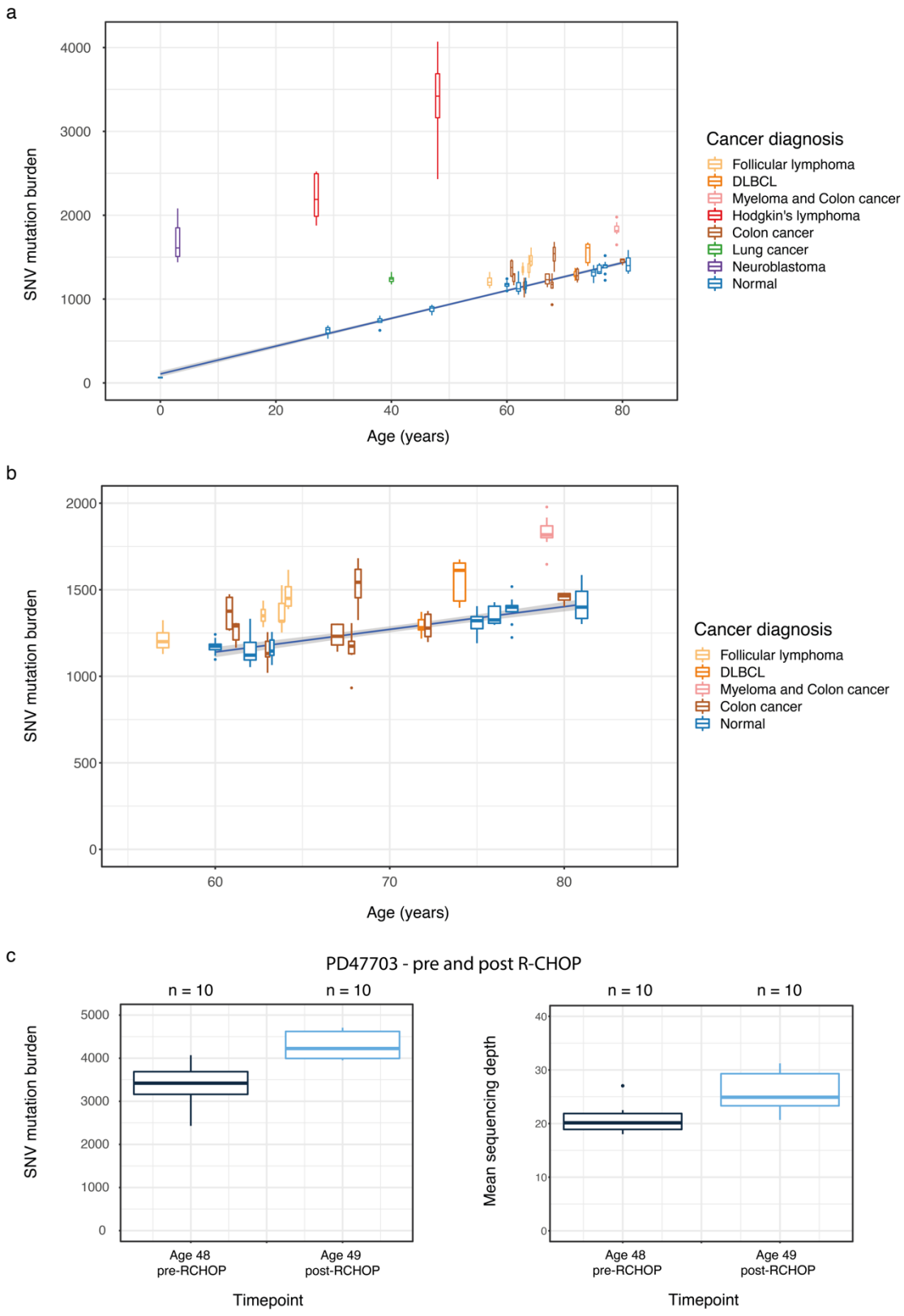


Fig. 5.2| SNV mutation burden in normal vs chemotherapy exposed HSPCs. a, Burden of single nucleotide variants across the normal and chemotherapy exposed cohort (n = 340) coloured by cancer diagnosis. The boxes indicate the median and interquartile range and the whiskers denote the range. The blue line represents a regression of age on mutation burden for the normal samples only, with 95% CI shaded. **b**, Plot as in **a** with data from the four individuals with the greatest burden

of excess mutations removed to allow better visualisation of the data from other individuals. **c**, Burden of single nucleotide variants pre and post R-CHOP chemotherapy for individual PD47703 (left). The samples were taken one year apart. The right-hand plot shows the mean sequencing depth for the same pre and post R-CHOP samples from individual PD47703.

Of the other patients, HSPCs from the individual with lung cancer had approximately 500 excess mutations. HSPCs from all but one of the other individuals treated for haematological malignancy (follicular lymphoma, DLBCL and myeloma) had 200-400 excess mutations. In contrast, in the bowel cancer cohort only 3 out of 8 individuals had HSPCs with a similar mild excess of 200-400 mutations. The remaining 5 individuals had HSPCs with mutation burdens in the normal range.

The results do suggest some inter-individual variability in mutation acquisition in normal HSPCs when treated with identical chemotherapeutic regimens. For example, of the two individuals treated with R-CHOP for DLBCL, one had a completely normal HSPC mutation burden, while the other harboured a few hundred excess mutations. In addition, the most heavily treated colon cancer patient, who received more than double the number of cycles of oxaliplatin (22 cycles) and 5FU/Cap (53 cycles) when compared to other patients in the cohort, showed no elevation in HSPC mutation burden. Nevertheless, the two colon cancer patients with the largest elevation in HSPC mutation burden were relatively heavily treated, having received 8 and 11 cycles of oxaliplatin and 14 and 16 cycles of 5FU/Cap respectively.

Although the data demonstrate an elevated mutation burden in many chemotherapy exposed individuals, the cause of the elevation based solely on these results remains unclear. Due to the fact that the majority of individuals have been treated with multiple agents, it is near impossible to link specific agents with the mutagenic insult. There is only one individual treated with a single agent (the alkylating agent bendamustine), who has an elevated mutation burden of around 300, potentially implicating this specific agent. It is also possible that chemotherapy induces mutation accumulation purely as a result of increased cell division. However it is unlikely this alone could explain the highest mutation burden results observed here, as recent work suggests that approximately just a single SNV is accumulated per cell division[3,86,87].

In further support for chemotherapy having a direct role on mutation accumulation, mutation burden data is available before and after R-CHOP chemotherapy for one individual (PD47703), who had previously received ChlVPP and BGemV. **Figure 5.2c** shows that the SNV mutation burden post-RCHOP was over 500 SNVs/cell greater than pre-RCHOP, despite the samples only being taken a year apart. Although the mean sequencing depth of the post-RCHOP samples was higher, this would only account for a maximum difference of around 100 mutations.

Indel burden in chemotherapy exposed individuals

Indel burden is more variable in normal individuals (**Fig. 3.2c**) and is over an order of magnitude lower than SNV burdens in HSPCs, meaning it is more difficult to detect small changes in indel mutation burden that may result from chemotherapeutic exposure. In contrast to the SNV mutation burdens, the majority of chemotherapy exposed individuals do not have an elevated indel burden compared to normal (**Fig. 5.3**). As perhaps expected, the notable exceptions are the four individuals with the greatest elevation in HSPC SNV burden, who show a similar degree of increase in their indel burdens (approximately 50-100 excess indels).

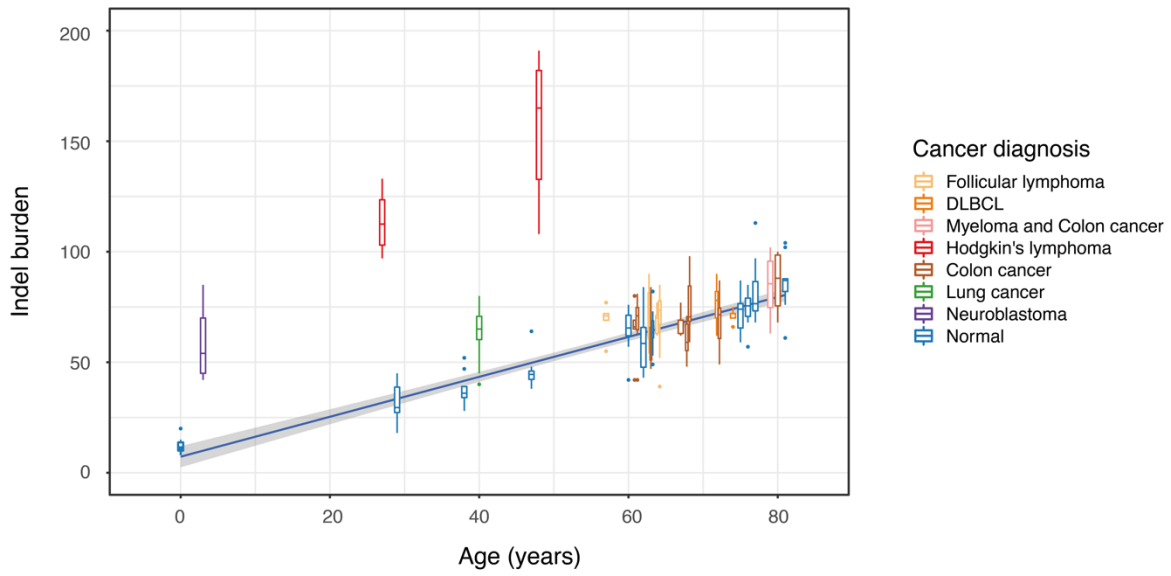


Fig. 5.3| Indel mutation burden in normal vs chemotherapy exposed HSPCs. Burden of indels across the normal and chemotherapy exposed cohort (n = 340) coloured by cancer diagnosis. The boxes indicate the median and interquartile range and the whiskers denote the range. The blue line

represents a regression of age on mutation burden for the normal samples only, with 95% CI shaded.

Mutation burden summary

1. HSPC SNV burden was highly variable across the chemotherapy exposed cohort, ranging from completely normal to a mean excess of over 2000 mutations.
2. Inter-individual variation in mutation burden in response to the same chemotherapeutic exposures was observed.
3. Individuals with markedly elevated HSPC SNV burdens, also had much higher variance in these burdens than observed in normal individuals, suggesting some variability in mutation accumulation between cells within an individual. Although in these high burden cases there were no cells identified with a completely normal mutation burden.
4. Indel burden was only demonstrably elevated in the four individuals with the greatest elevation in SNV burden.
5. Which specific chemotherapeutic agents are implicated in the elevated HSPC mutation burdens remains unclear from this data alone.

5.3 Mutational signatures of chemotherapy

Mutational signature extraction was performed *de novo* on a total of 435 samples in the study using a Hierarchical Dirichlet Process (HDP, https://github.com/nicolaroberts/hdp; **Chapter 2**). The HDP analysis included all samples in the 'mutation burden' cohort plus 35 additional samples per individual from the chemotherapy exposed 'population structure cohort' to generate phylogenies for these five individuals of 40-45 samples in size. The larger phylogenies allow better resolution of timing and cell-to-cell variation of mutagenic insult, which is particularly important for the individual who received chemotherapy both as a child in addition to more recently.

In total eight signatures were extracted from the data (**Fig. 5.4**). These included three signatures found in normal samples (**Fig. 5.5**). Components 1 and 2 are very similar to the previously well described COSMIC signatures SBS1 and SBS5 (age-related processes).

Component 3 represents 'SBS-Blood', which has been previously reported in normal blood[53,86]. Five signatures (Components 4, 5, 6, 7 and 8) were found exclusively in the chemotherapy exposed cohort. The pattern of individuals in whom the signatures were found (**Figs. 5.6, 5.7, 5.8 and 5.9**) allowed us to determine the specific mutagenic chemotherapy agents responsible, as discussed below.

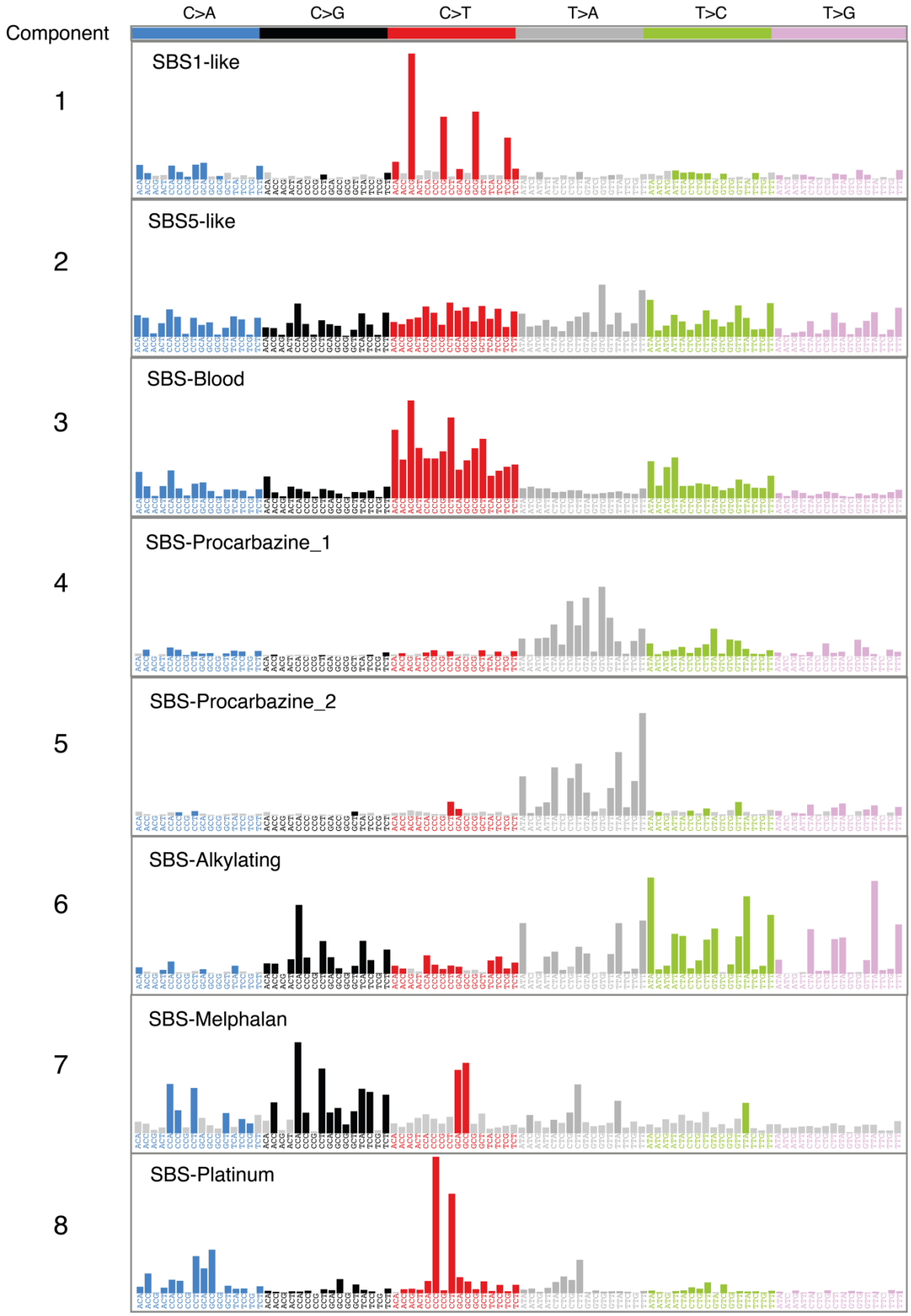


Fig. 5.4| HDP extracted mutational signatures (run with no priors). Barplots showing the trinucleotide distribution of the HDP-derived signatures extracted from the dataset. A total of 8 *de novo* signatures were extracted.

Mutational signatures found in normal individuals

Three mutational signatures were found in normal individuals (Components 1, 2 and 3). Components 1 and 2 are very similar to Cosmic signatures SBS1 and SBS5, known to be due to age-related processes (**Fig. 5.5**). The predominant mutational process in all the normal adults was Component 3 (SBS-Blood), a signature that has been previously reported in normal HSPCs[53,86]. In contrast the predominant mutational process in the cord blood samples was SBS1-like (data not shown), as previously found in other cord blood and foetal HSPCs[3,53]. SBS1-like was present at significant levels in two adult individuals, PD40667 (individual with inflammatory bowel disease) and PD41048 (individual with selenoprotein deficiency). SBS5-like was present at low levels in the two oldest individuals, and interestingly timed to the later branches of expanded clades, potentially suggesting that replicative stress (or another age-related process) may be the underlying aetiology.

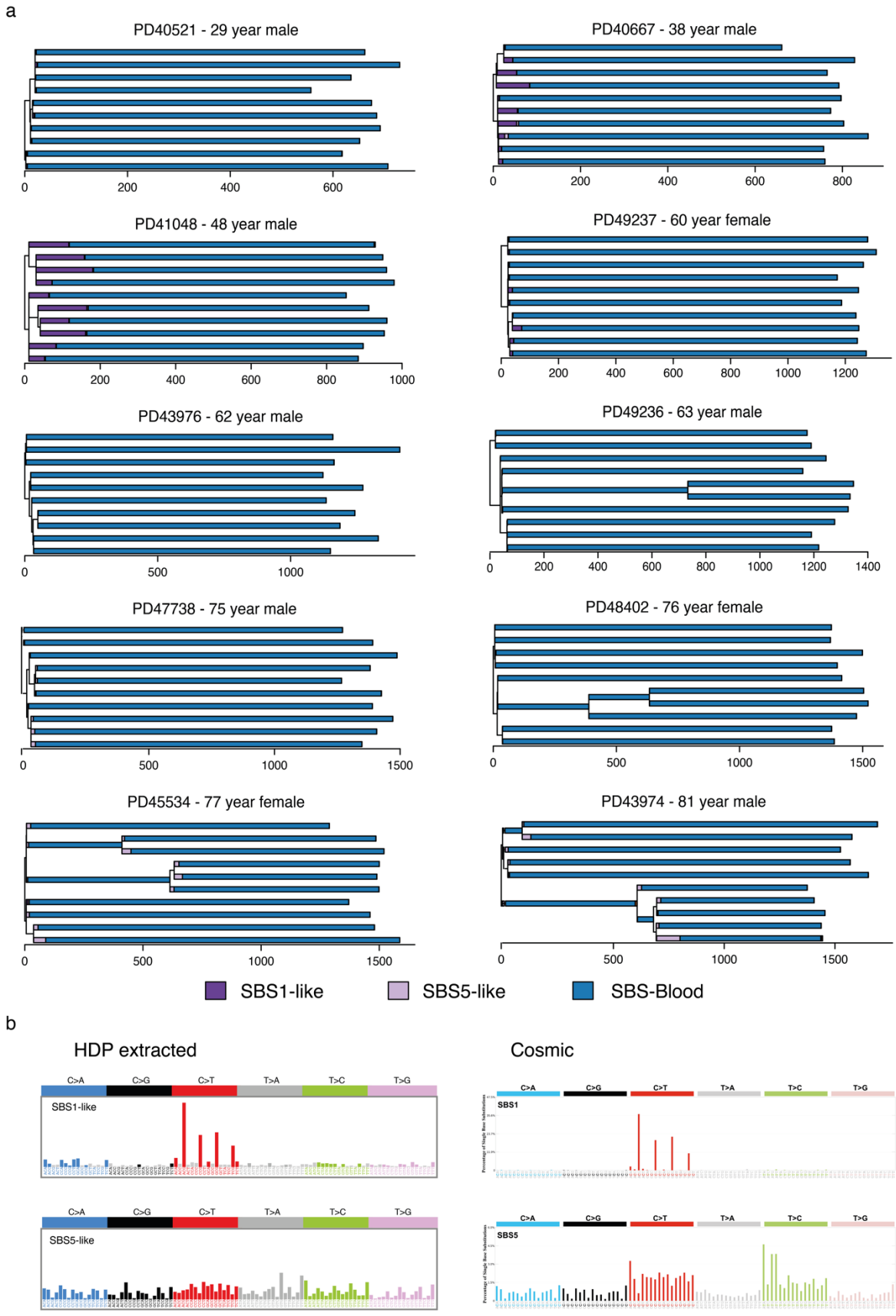


Fig. 5.5| Mutational signatures in normal individuals. a, Phylogenetic trees reconstructed from somatic mutations for all the normal adult individuals in the cohort. Branch lengths are proportional

to the number of somatic mutations. The axes show mutation number. Bar plots are overlaid on each phylogenetic branch to illustrate the relative contributions of each signature. **b**, A comparison of the *de novo* HDP-extracted signatures SBS1-like and SBS5-like, and SBS1 and SBS5 as depicted in the COSMIC database.

Novel nitrogen mustard alkylating agent signature

A novel signature (Component 6) was identified in individuals exposed to the nitrogen mustard alkylating agents bendamustine (PD47541 and PD47703), and chlorambucil (PD44703) (**Fig. 5.6**). In these two individuals the alkylating agent signature conferred between 200-1500 additional mutations. The data presented in **Fig. 5.6a** allows us to confidently attribute Component 6 to nitrogen mustard alkylating agents as PD47541, exposed only to bendamustine, has approximately 300 excess mutations per cell attributed to this signature, so causally implicating this single agent (rather than potentially a combination of agents).

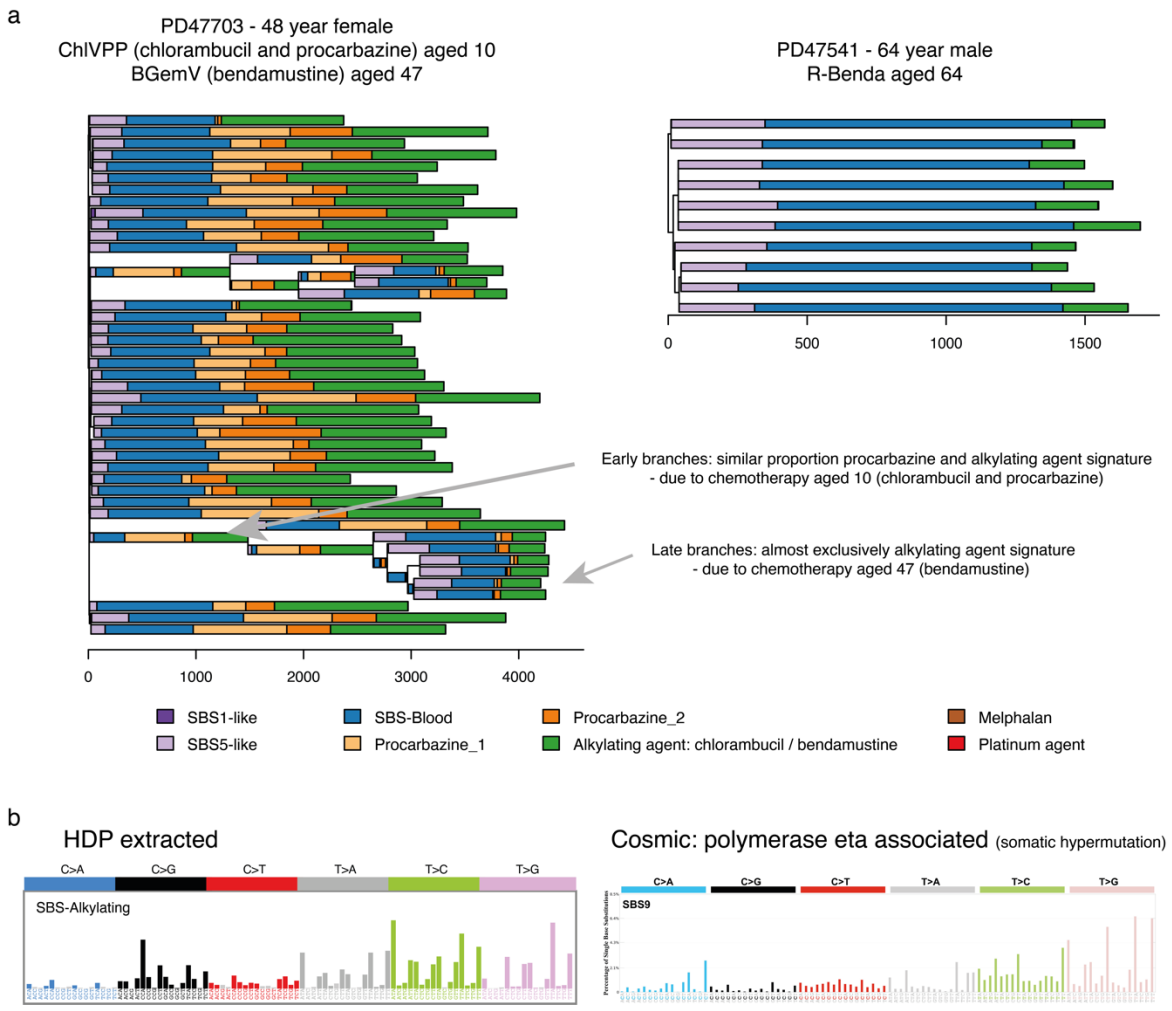


Fig. 5.6| Mutational signatures in individuals exposed to chlorambucil and bendamustine. a, Phylogenetic trees reconstructed from somatic mutations for the two individuals with SBS-Alkylating. Branch lengths are proportional to the number of somatic mutations. The axes show mutation number. Bar plots are overlaid on each phylogenetic branch to illustrate the relative contributions of each signature. **b**, A comparison of the *de novo* HDP-extracted signatures SBS-Alkylating, and SBS9 as depicted in the COSMIC database.

PD47703 received two different bifunctional alkylating agents: chlorambucil aged 10 and bendamustine aged 47. The clonal structure of her phylogeny allows timing of the 'nitrogen mustard alkylating agent signature' to both early and late time points, indicating that both chlorambucil and bendamustine produce the same signature. However, there is approximately double the amount of alkylating agent signature in the early timepoint branches compared to the late timepoint, which may be due to chlorambucil having greater mutagenic toxicity, or to the regimen ChlVPP being more intensive than BGemV. The impact

of bendamustine therapy in terms of numbers of mutations conferred is similar between PD47703 (6 cycles BGemV) and PD47541 (5 cycles R-Benda).

The nitrogen mustard derivative alkylating agents are described as 'bifunctional', possessing two alkylating groups on a single molecule. Due to this, they can form covalent bonds at two nucleophilic sites on different DNA bases, resulting in the formation of both intrastrand and interstrand DNA crosslinks. The observed signature in individuals exposed to bifunctional alkylating agents shows a pattern of peaks at TpW (TpT and TpA), and is similar to SBS9, a signature associated with somatic hypermutation in B cells (**Fig. 5.6b**). Polymerase eta (a translesional polymerase) has been causally implicated in B cell somatic hypermutation[192] and translesional polymerases are also known to be involved in repair of DNA lesions including DNA crosslinks[193,194]. Translesional polymerases possess reduced replicative fidelity due to a large active site that can accommodate DNA adducts. This is of benefit in allowing them to replicate past DNA lesions, but characteristic patterns of their infidelity are also the likely cause of the mutational signature we observe in the individuals exposed to bifunctional alkylating agents.

However, the signature was not identifiable in individuals exposed to another bifunctional nitrogen mustard alkylating agent, cyclophosphamide (**Fig. 5.7**). The relative lack of mutagenesis observed with cyclophosphamide as compared to the other bifunctional alkylating agents can also potentially be explained at the molecular level. Cyclophosphamide is thought to relatively spare stem and progenitor cells because these cells possess high levels of ALDH1. The ALDH1 enzyme results in the inactivation of aldophosphamide (an intermediate metabolite) prior to its conversion to the toxic metabolites phosphoramide mustard and acrolein, resulting in reduced stem and progenitor cell toxicity. Our data would support this view.

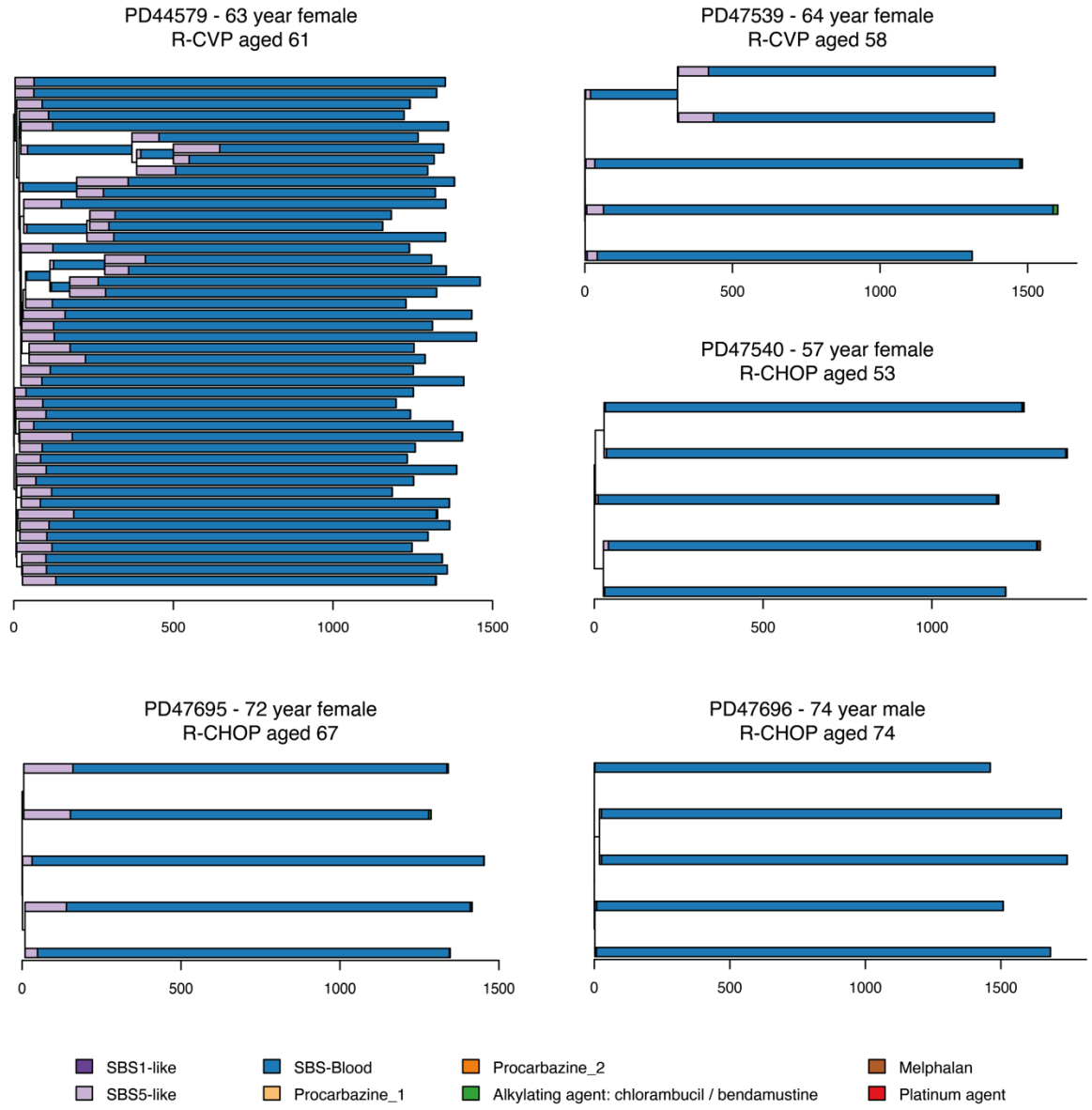


Fig. 5.7| Mutational signatures in individuals exposed to cyclophosphamide. Phylogenetic trees reconstructed from somatic mutations for five individuals exposed to cyclophosphamide. Branch lengths are proportional to the number of somatic mutations. The axes show mutation number. Bar plots are overlaid on each phylogenetic branch to illustrate the relative contributions of each signature.

Monofunctional alkylating agent signature

Two signatures (Components 4 and 5) were only identified in two individuals (PD47703 and PD50308) who were both treated for Hodgkin's lymphoma (**Fig. 5.8a**). PD47703 was first diagnosed with Hodgkin's lymphoma aged 10, at which time she received ChlVPP (chlorambucil, vinblastine, procarbazine and prednisolone). She was subsequently diagnosed with Hodgkin's lymphoma again aged 47, at which time she was treated with BGemV

(bendamustine, gemcitabine and vinorelbine). Her blood was sampled 8 months post BGemV aged 48. PD50308 was diagnosed with Hodgkin's lymphoma aged 25 for which she received the regimen escBEACOPP (bleomycin, etoposide, doxorubicin, cyclophosphamide, vincristine, procarbazine and prednisolone). Her blood was sampled 2 years later aged 27. The only drug that both these individuals were exposed to was procarbazine, a monofunctional alkylating agent (single reactive site), implicating this as the cause of both these T>A signatures. In PD47703 the structure of the phylogeny allows timing of the putative 'procarbazine signature' early in life, which would fit with it having arisen during her treatment with procarbazine aged 10. Some elements of SBS-Procarbazine_2 overlap with the 'Alkylating agent' signature, explaining why this is less dominant in PD50308, who received cyclophosphamide, which does not have a mutagenic impact in normal HSPCs, in contrast to chlorambucil.

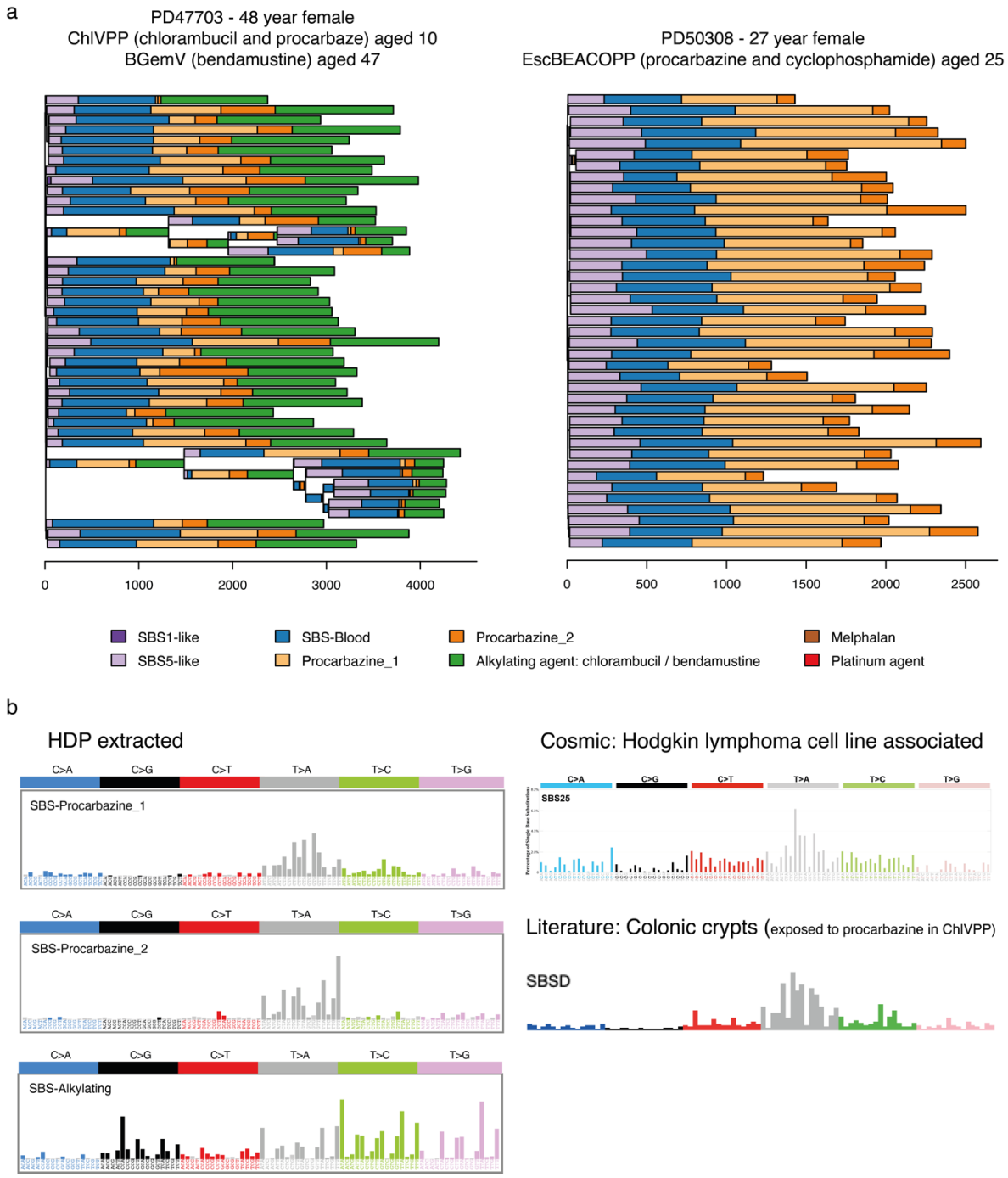


Fig. 5.8| Mutational signatures in individuals exposed to procarbazine. a, Phylogenetic trees reconstructed from somatic mutations for the two individuals exposed to procarbazine. Branch lengths are proportional to the number of somatic mutations. The axes show mutation number. Bar plots are overlaid on each phylogenetic branch to illustrate the relative contributions of each signature. **b**, A comparison of the *de novo* HDP-extracted signatures SBS-Procarbazine_1, SBS-Procarbazine_2 and SBS-Alkylating, with SBS25 as depicted in the COSMIC database and SBSD from literature[101].

COSMIC signature SBS25 is very similar to the SBS-Procarbazine_1 signature reported here. SBS25 has previously only been identified in Hodgkin's lymphoma cell lines. Treatment information is available for one of the SBS25 containing Hodgkin's lymphoma cell lines, which was derived from tumour tissue known to have been exposed to procarbazine in the regimen COPP/ABVD[195]. A similar signature has also previously been identified in normal colonic crypts from an individual treated for Hodgkin's lymphoma (cosine similarity to SBS25 = 0.9)[101] (**Fig. 5.8b**). The individual also received procarbazine in the regimen ChlVPP/PABIOE, providing further evidence that this is the causal agent for Component 4.

Procarbazine was the only monofunctional alkylating agent included in our cohort, but some evidence suggests that dacarbazine, the other commonly used monofunctional alkylating agent used in Hodgkin's lymphoma regimens may be less toxic than procarbazine to normal tissues. Procarbazine is much more likely than dacarbazine to lead to infertility in both males and females, and was removed from paediatric regimens in the UK some years ago for this reason. An ongoing UK based clinical study of adult therapy with escBEACOPDac (where procarbazine has been replaced with dacarbazine) has found significantly reduced blood transfusion requirement for patients undergoing therapy (unpublished data). As another line of evidence, a father who had previously received chemotherapy including both iphosphamide (another nitrogen mustard alkylating agent) and dacarbazine, passed on a similar signature to the 'nitrogen mustard alkylating agent signature' we describe to his offspring's germline, but there was no evidence of an SBS25-like signature (or any other additional signature)[196].

Melphalan signature

Melphalan is another nitrogen mustard derivative alkylating agent, most commonly used in the treatment of myeloma. HDP extracted a signature (Component 7), unique to PD47699 who had received melphalan for myeloma (**Fig. 5.9a**). This was similar to SBS-MM1 reported in the myeloma cells of individuals who relapsed after melphalan autograft[97,197,198], suggesting that Component 7 was most likely due to melphalan therapy and therefore referred to as SBS-Melphalan here. (**Fig. 5.9b**).

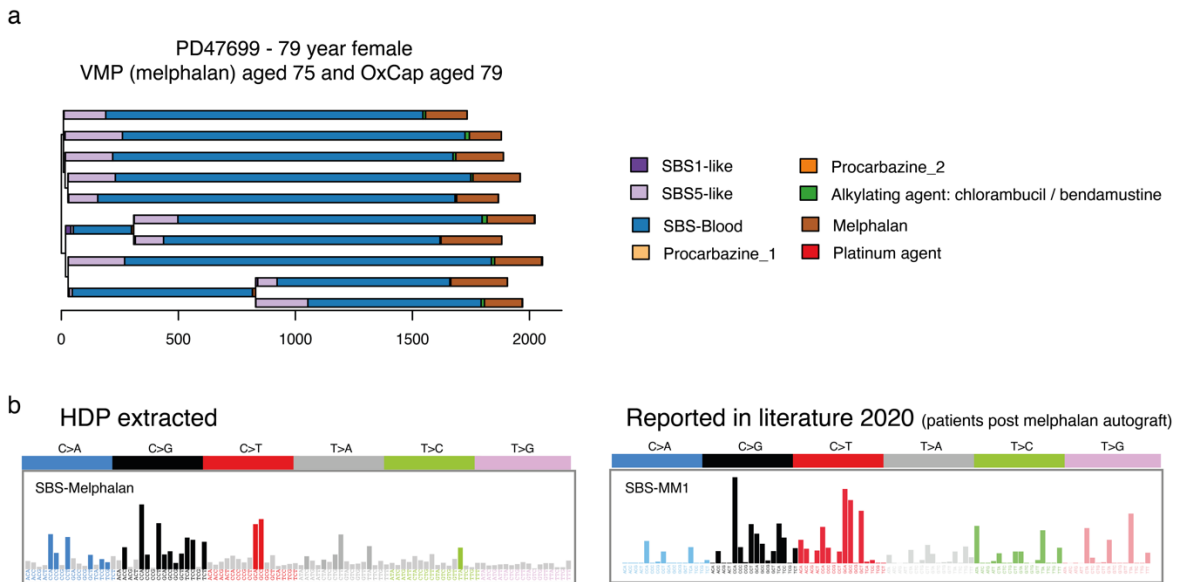


Fig. 5.9| Mutational signatures in individual exposed to melphalan. a, Phylogenetic trees reconstructed from somatic mutations for the one individual exposed to melphalan. Branch lengths are proportional to the number of somatic mutations. The axes show mutation number. Bar plots are overlaid on each phylogenetic branch to illustrate the relative contributions of each signature. **b**, A comparison of the *de novo* HDP-extracted signatures SBS-Melphalan, and SBS-MM1 (previously ascribed to melphalan) as depicted in recent literature[97,197,198].

Platinum agent signature

The final signature extracted in the cohort of chemotherapy-exposed individuals was Component 8 or SBS-Platinum. SBS-Platinum is almost identical to the COSMIC signature SBS31, with some additional contribution from COSMIC signature SBS35, both known to be related to exposure to cisplatin/carboplatin. This signature contributed the majority of excess mutations in two patients, PD50306 and PD50307 (**Fig. 5.10a**). PD50306 received cisplatin and carboplatin, in the regimen rapidCOJEC for neuroblastoma and PD50307 received a combination of carboplatin and etoposide for lung cancer. SBS-Platinum was the only chemotherapy-associated mutational signature present in these individuals.

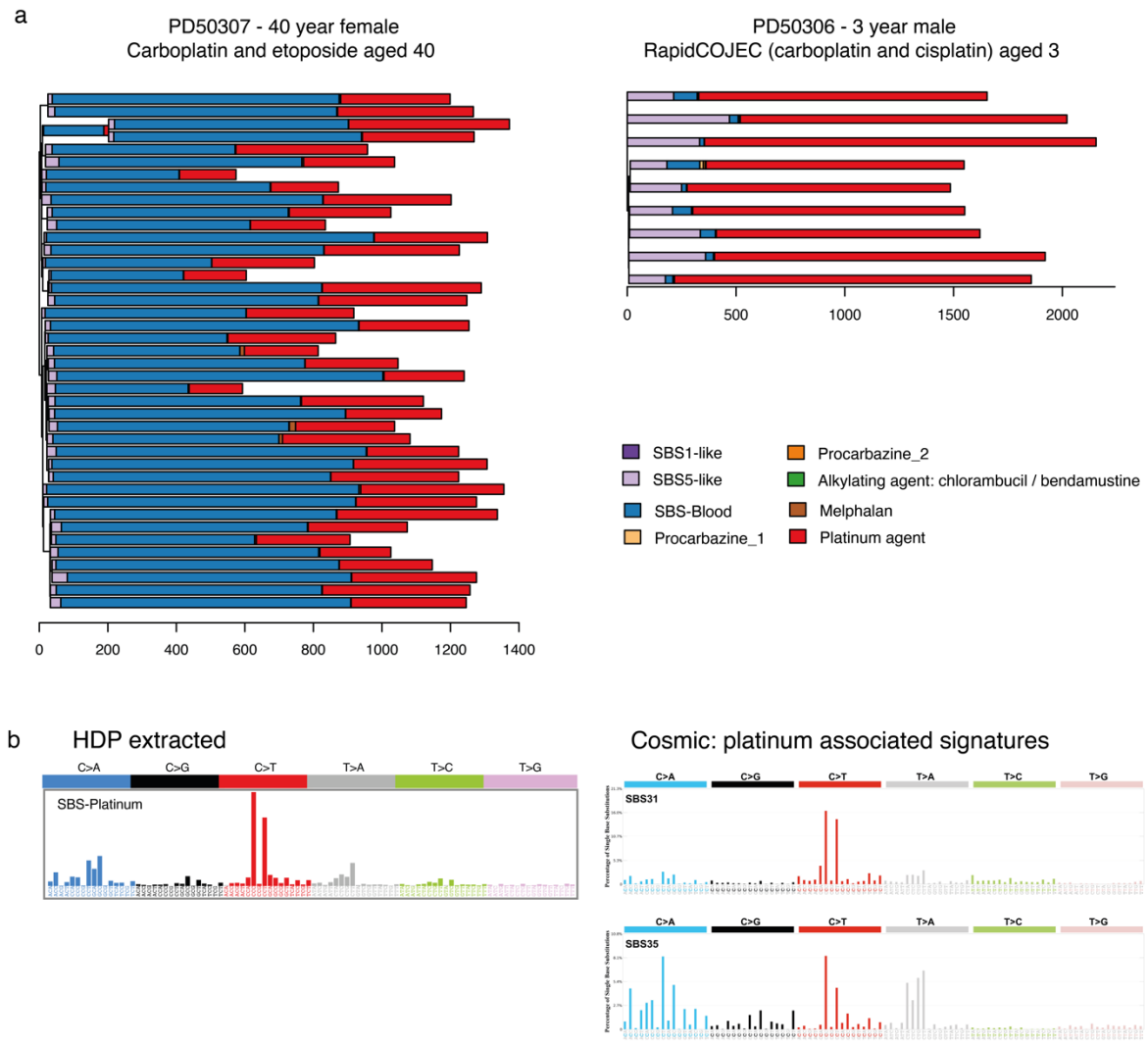


Fig. 5.10| Mutational signatures in individuals exposed to carboplatin and or cisplatin. a, Phylogenetic trees reconstructed from somatic mutations for the two individuals exposed to cisplatin and / or carboplatin. Branch lengths are proportional to the number of somatic mutations. The axes show mutation number. Bar plots are overlaid on each phylogenetic branch to illustrate the relative contributions of each signature. **b**, A comparison of the *de novo* HDP-extracted signatures SBS-Platinum, and SBS31 and SBS35 as depicted in the COSMIC database.

SBS31 was originally identified in cisplatin exposed human cell lines and oesophageal and head and neck cancers of patients treated with cisplatin prior to surgery[199] (**Fig. 5.10b**). More recently SBS31 has been identified in two cases of paediatric therapy-related myeloid neoplasm (TMN) occurring within a year of completing therapy for high-risk neuroblastoma[102]. Both TMNs had over 2000 mutations, similar to the burden we observed

in all the normal HSPCs in PD50306. SBS31 has also been identified in a large study of metastatic cancer samples[99].

In a parallel to our observations with nitrogen mustard alkylating agents, while we identified a striking signature and elevation in mutation burden in individuals treated with cisplatin and / or carboplatin, we did not observe such striking evidence of chemotherapy induced mutagenesis in the six individuals exposed to oxaliplatin (**Fig. 5.11a**). This was despite three individuals in the cohort receiving 10 or more cycles of oxaliplatin. Only 3 out of 7 exposed individuals showed very low levels of the 'platinum signature' (PD47537, PD47701 and PD47700).

The study of metastatic cancer samples mentioned above[99] did identify a novel 'oxaliplatin' signature at high burden in metastatic colon cancers that we have not extracted independently in the normal blood cells of our cohort (**Fig. 5.11b**). To my knowledge this is the only study to have identified a mutational signature specifically relating to oxaliplatin. The 'oxaliplatin' signature does however have some similarities to SBS-Platinum.

The reduced impact of oxaliplatin on normal HSPCs fits with what is known about the toxicities of the platinum agents. Carboplatin is the most myelotoxic of the three agents, while oxaliplatin has very little impact on peripheral blood counts[200,201]. In addition, oxaliplatin is known to form fewer DNA crosslinks at therapeutic doses, and is thought to have additional, as yet unclear, anti-cancer mechanisms of action.

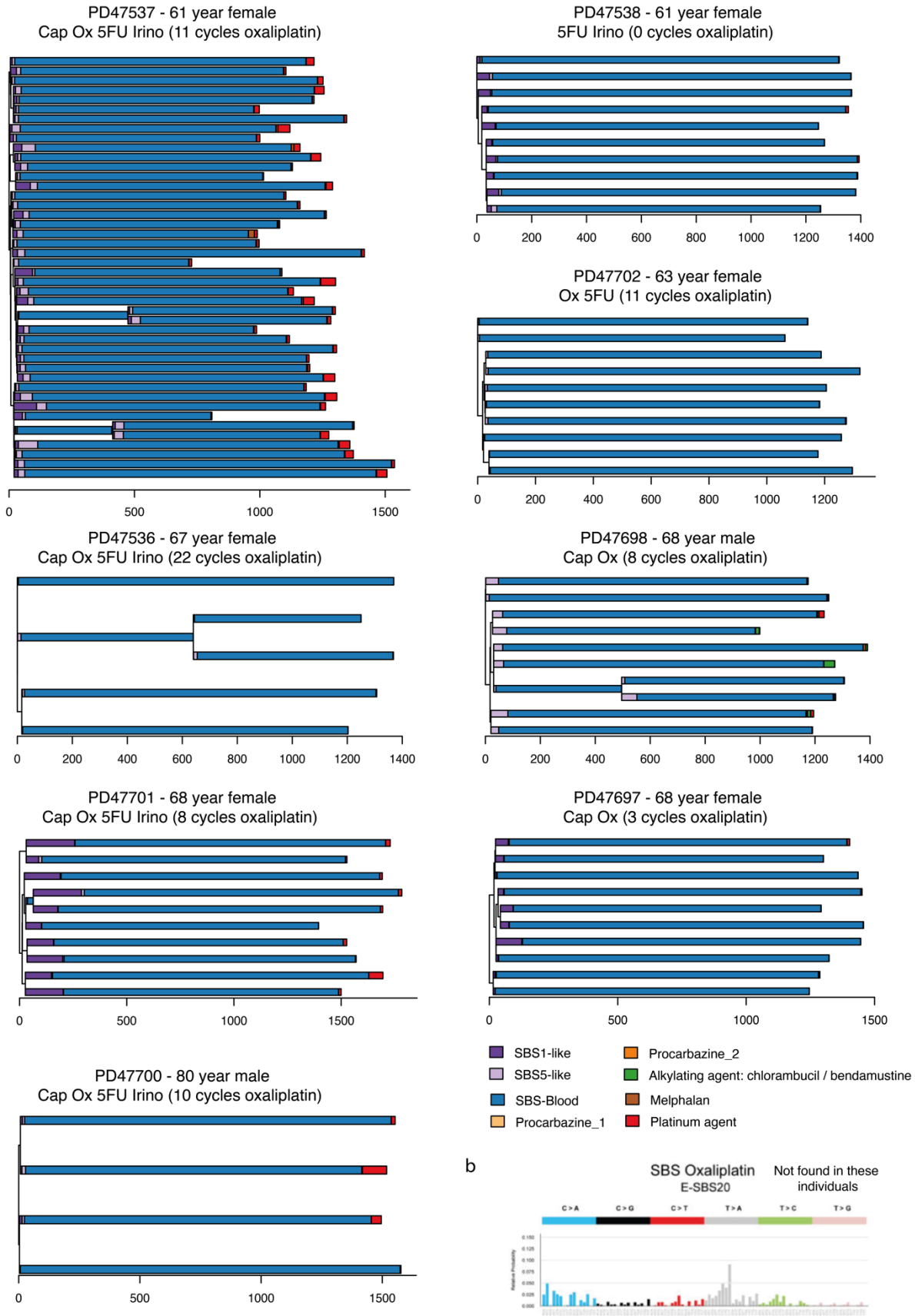


Fig. 5.11| Mutational signatures in individuals exposed to oxaliplatin. a, Phylogenetic trees reconstructed from somatic mutations for the two individuals with SBS-Alkylating. Branch lengths

are proportional to the number of somatic mutations. The axes show mutation number. Bar plots are overlaid on each phylogenetic branch to illustrate the relative contributions of each signature. **b**, SBS-Oxaliplatin as depicted in a recent study[99].

Other agents: topoisomerase inhibitors, anti-metabolites and vinca-alkaloids

We found no evidence of SBS mutational signatures in normal HSPCs relating to the topoisomerase I inhibitor irinotecan, the topoisomerase II inhibitors etoposide and doxorubicin, the anti-metabolites gemcitabine, 5FU and capecitabine, the vinca-alkaloids vincristine, vinblastine and vinorelbine or the cytotoxic antibiotic bleomycin.

In keeping with our findings, a 5FU/Capecitabine signature (SBS17-like) has been identified in colonic cancers and normal colonic crypts closely adjacent to the cancer, but not in other more spatially distant normal crypts[100]. Structural variant analysis will be more informative in assessing the impact of the topoisomerases, with topo II inhibitors in particular implicated in causing secondary myeloid malignancies[202].

Age-related signatures SBS1 and 5 in chemotherapy-exposed individuals

The normal 'age-related' signature SBS5-like (Component 2) was found to be markedly elevated in many individuals treated with chemotherapy, particularly those who received the most intensive chemotherapy regimens (PD50306, PD47703 and PD50308). It was also found at relatively high levels in individuals that received less intensive therapies, and may therefore explain the elevated mutation burdens observed in individuals who received RCVP and RCHOP, in the absence of evidence for a mutagenic impact of chemotherapy. In all individuals the SBS5-like component times to late branches in the phylogeny, fitting with it being related to chemotherapy. This potentially implicates increased cell turnover as the aetiology for SBS5, meaning SBS5 rather than SBS1 is the more likely of these two age-related mutagenic processes to be related in some way to cell division.

Mutational signature summary

1. A novel SBS signature relating to the nitrogen mustard derivative alkylating agents bendamustine, melphalan and chlorambucil was identified.

2. The nitrogen mustard derivative alkylating agent SBS signature is not present in individuals exposed to cyclophosphamide, despite it being a member of this class.
3. Two SBS25-like signatures, previously only identified in Hodgkin's lymphoma cell lines, are attributable to procarbazine.
4. An SBS31-like platinum signature was identified in the HSPCs of individuals exposed to cisplatin and carboplatin.
5. The SBS31-like platinum signature was present at low level in a minority of samples exposed to oxaliplatin. No other oxaliplatin specific signature was identified.
6. No SBS signatures can be attributed to the topoisomerase inhibitors, anti-metabolites, vinca-alkaloids or cytotoxic antibiotics included in this study.
7. SBS5 is elevated in many individuals exposed to chemotherapy, particularly prominent for the most intensive regimens (Hodgkin's lymphoma and neuroblastoma).

5.4 Population structure in chemotherapy exposed individuals

Evidence for changes in population size

The extent to which different chemotherapy regimens impact the population size of human HSCs remains an unanswered and interesting question. Murine studies have shown that LT-HSC populations do not recover to a normal size post depletion, typically remaining at a ceiling of 10% of the normal population size[103]. In humans, reduced resilience of the haematopoietic system is observed in individuals previously heavily treated with cytotoxic agents. With successive intensive regimens, even in younger individuals bone marrow recovery frequently becomes markedly slower and dose reductions are more frequently required[203,204]. However, it is not clear if these observations are due to a reduced HSC population or a more 'aged' or poorly functioning HSC population due to historic replicative stress. This reduced resilience of the haematopoietic system was observed clinically in PD47703, who received ChlVPP chemotherapy aged 10. She developed prolonged cytopenias during R-CHOP chemotherapy aged 48, requiring both cycle delay and dose reduction.

The larger chemotherapy-exposed phylogenies that give us the best potential insight into population size changes are shown along with age and size-matched normal phylogenies in **Figures 5.12 and 5.13**.

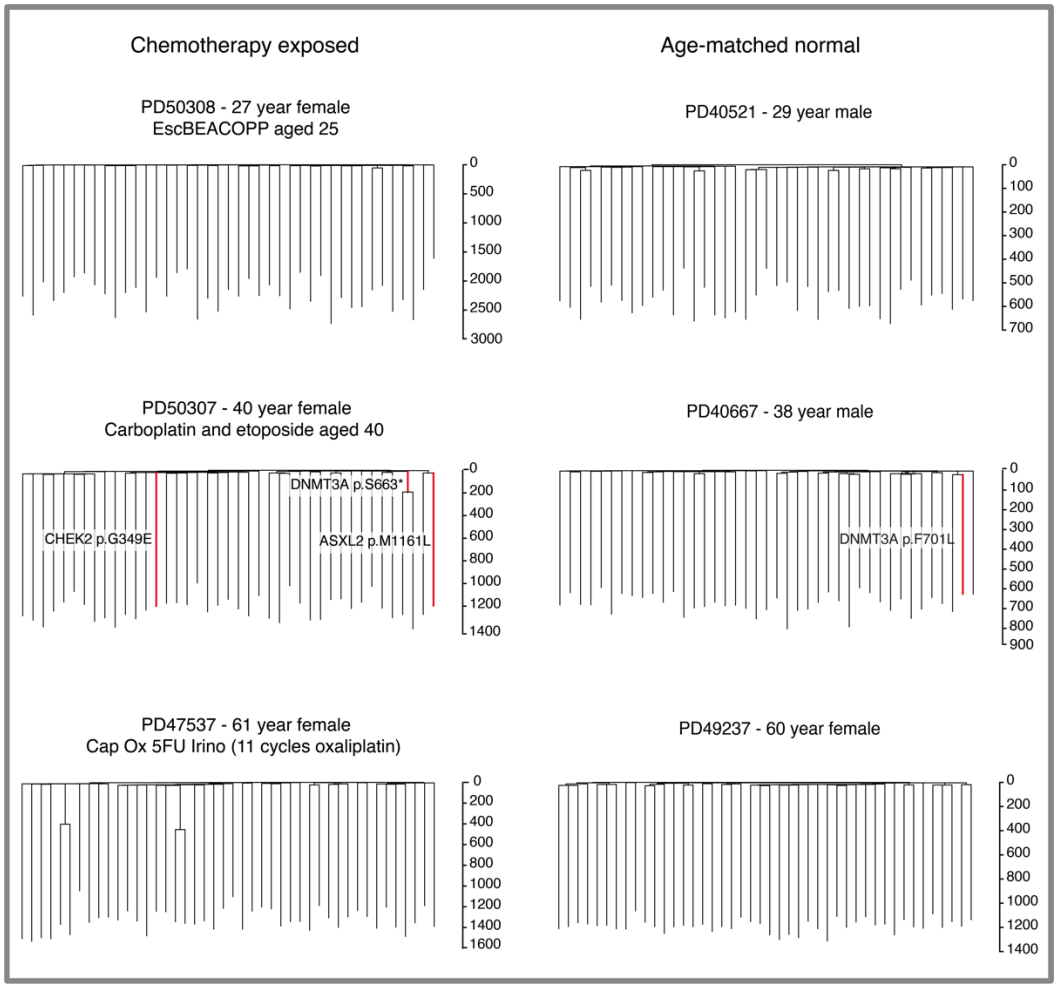


Fig. 5.12| Comparison of phylogeny structure and driver mutation burden: small phylogenies. Phylogenetic trees reconstructed from somatic mutations for three chemotherapy-exposed individuals (left), with age and size-matched normal phylogenies alongside (right). Axes show mutation number. Branch lengths are adjusted for mean sequencing depth. Probable and definite pathogenic mutations in any one of the top 17 clonal haematopoiesis genes (**Table 3.2**) are depicted, as well as mutations in CHEK2 which have been found to be associated with previous chemotherapy[106].

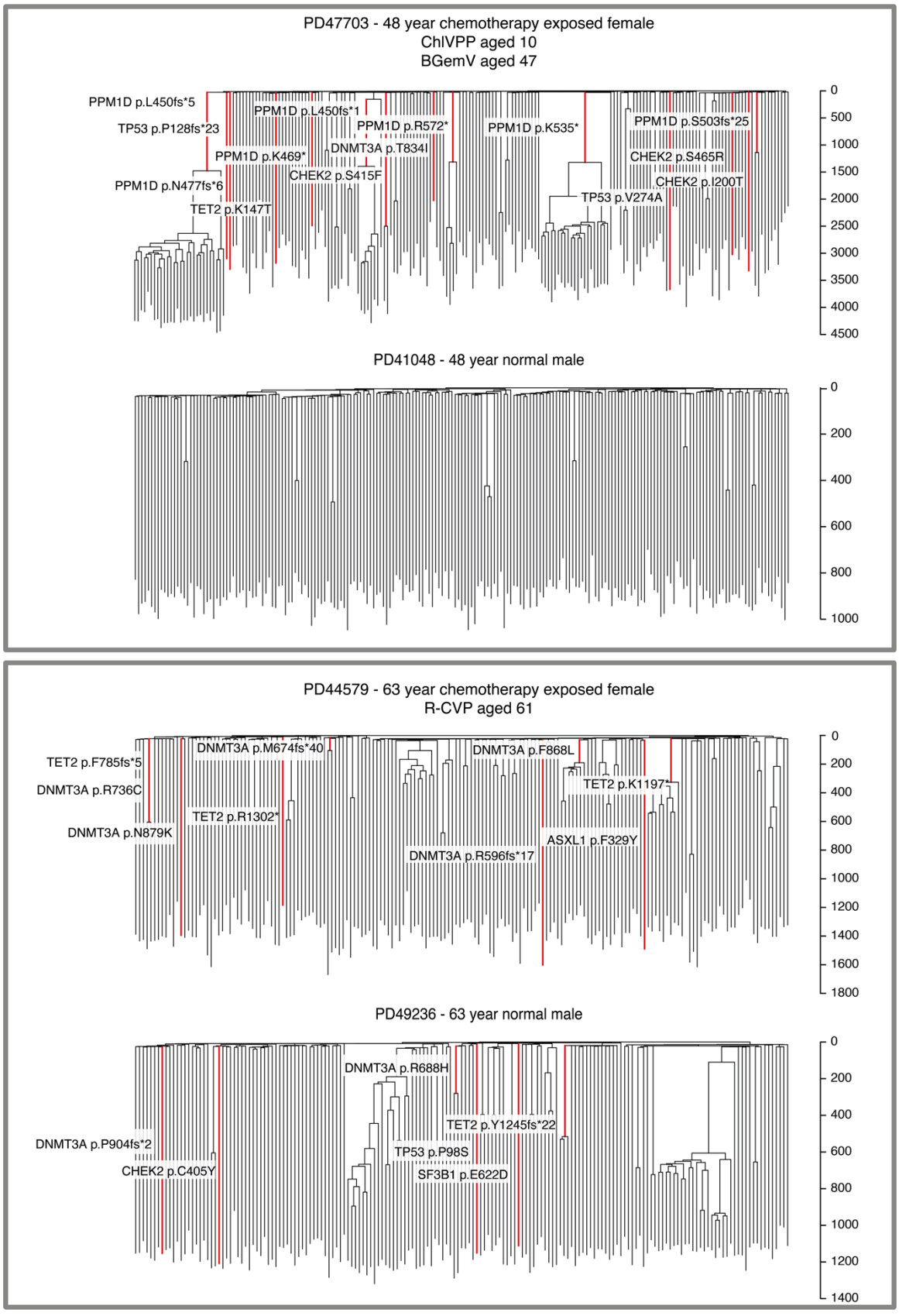


Fig. 5.13| Comparison of phylogeny structure and driver mutation burden: large phylogenies. Phylogenetic trees reconstructed from somatic mutations for two chemotherapy-exposed individuals, with age and size-matched normal phylogenies below. Axes show mutation number.

Branch lengths are adjusted for mean sequencing depth. Probable and definite pathogenic mutations in any one of the top 17 clonal haematopoiesis genes (**Table 3.2**) are depicted, as well as mutations in CHEK2 which have been found to be associated with previous chemotherapy[106].

As discussed in **Chapter 4**, LT-HSC population size changes can only be accurately assessed in the absence of positive selection and in the time window before ST-HSC/MPPs are contributing (>15 years prior to the time of sampling). In the few years before sampling, when the majority of these individuals received chemotherapy, there is an additional contribution from ST-HSC/MPPs to total population size. This means that only very severe reductions in the population size of this whole compartment could be visualised (for example from > 1 million to < 100,000). Large phylogenies (>300 samples) also enable more accurate population size inferences due to better resolution at higher population sizes. For these reasons the relatively small phylogenies presented here are suboptimal for making a good assessment of population size changes as a result of chemotherapy. However, despite these clear limitations, some broad conclusions can be made.

The best phylogeny for making population size inferences is PD47703, who received chemotherapy aged 10 and whose HSC population was sampled 38 years later. This phylogeny has a number of clades under positive selection due to *PPM1D* mutations, but outside the expanded clades the population remains relatively polyclonal. From this pattern we can be sure there was no catastrophic decrease (> 10-fold) in LT-HSC population size resulting even from intensive chemotherapy such as ChlVPP.

The other phylogenies are all from individuals in whom chemotherapy was given in the 5 years prior to samplinig, making it more difficult to make any firm conclusions regarding population size changes. However, there are no coalescences observed in that time window in any of the phylogenies, showing they are also consistent with there being no catastrophic decline in the HSC/MPP population as a result of chemotherapy in these individuals.

Evidence for changes in clonal dynamics

To look for evidence of changes in clonal dynamics, the chemotherapy exposed phylogenies in **Figures 5.12 and 5.13** can be compared to similar sized and age matched normal

phylogenies. The phylogeny for PD47703 provides the most striking evidence for a direct effect of chemotherapy altering clonal dynamics. The work on normal individuals in **Chapter 4** would predict very few if any expanded clades detectable in an individual aged 48 as shown by our age matched normal phylogeny. However, this individual has 4 expanded clades, all carrying *PPM1D* driver mutations, and an additional 3 *PPM1D* mutations and 1 *TP53* mutation in cells from non-expanded clades. *PPM1D*, *TP53* and *CHEK2* driven clonal haematopoiesis has been shown to be more prevalent in individuals who have had previous chemotherapy in keeping with our observations in this individual[106,107]. As a result of the *PPM1D* clonal expansions, the PD47703 phylogeny is markedly more oligoclonal than the age and size-matched normal, with a Shannon Diversity index of 7.6 compared to 17.6 for PD41048.

A second blood sample was taken from PD47703 a year after the first sample during which time she was treated with R-CHOP chemotherapy for DLBCL. A total of 58 HSC/MPP (Lin -, CD34+, CD38-, CD45RA-; rather than HSPC) derived colonies from this second timepoint underwent whole genome sequencing. Interestingly the efficiency of colony formation of these colonies was only 22%, much lower than had been observed in even the elderly individuals sampled for the ageing project (**Fig. 2.2a**). This raises the possibility that a fraction of the HSC/MPP population in this individual is senescent or exhausted and so unable to proliferate *in vitro*.

Between the sampled timepoints, and most likely as a result of the additional chemotherapy, three of the *PPM1D* clades approximately doubled in size, implying a fitness effect of approximately 100% over 1 year (**Fig. 5.14**). This is a much higher fitness effect than those observed in the largest clades of the normal elderly phylogenies. However, the fact that many of the *PPM1D* clades were still only relatively small in size at the time of first blood sampling (despite having originated at the time of first chemotherapy aged 10) shows this high fitness effect is most likely transient, occurring only during the months over which chemotherapy is given. This fits with the prediction that *PPM1D* mutations only confer a significant advantage during the period of chemotherapy administration. For many of the *PPM1D* mutations it is not clear from the phylogeny if the mutation was caused by the mutagenic insult of chemotherapy, or was pre-existing in the HSPC population. However at least one mutation, *PPM1D p.K535** can be timed to *in utero* in the phylogeny from the second timepoint.

One caveat to the calculation of *PPM1D* mutation fitness effects from the two phylogenies is the fact that the post-RCHOP phylogeny contains both HSC/MPPs and HPCs, while the pre-RCHOP phylogeny was solely HPC derived. An alternative explanation for the difference in clade sizes between the two phylogenies is therefore that *PPM1D* mutated cells preferentially expand within the stem cell compartment, but contribute fewer downstream progeny than non-mutated cells.

Another interesting observation in this individual is the presence of a co-occurring *CSF3R* loss of function mutation in one of the *PPM1D* expanded clades (**Fig. 5.14**). *CSF3R* encodes the receptor for colony-stimulating factor 3 and is thought to play an important role in the growth and differentiation of granulocytes. Loss of function mutations in *CSF3R* have been described in patients with severe congenital neutropenia and HSPCs carrying this mutation would be predicted to have severely reduced differentiation towards granulocytes. This therefore illustrates one of the themes of the previous chapter, namely that acquisition of positively selected somatic mutations can confer age-related loss of function phenotypes.

One overall caveat to the clonal dynamics findings presented here, is that driver population frequencies have only been assessed in progenitor cell compartments, and have not yet been assessed in the corresponding mature blood cell compartments of these individuals. This has been performed in other individuals (published[86] and unpublished work), in whom similar somatic variant allele fractions were found in progenitor derived trees compared to myeloid mature cell compartments. However, it remains possible, that some variants are less well represented in the mature cell compartments compared to the HPC compartment. Of note, compartmental skewing of somatic variant allele fractions is something we have observed between the HSC and HPC compartments in the elderly phylogenies, most marked in one specific clade (**Fig. 4.5**).

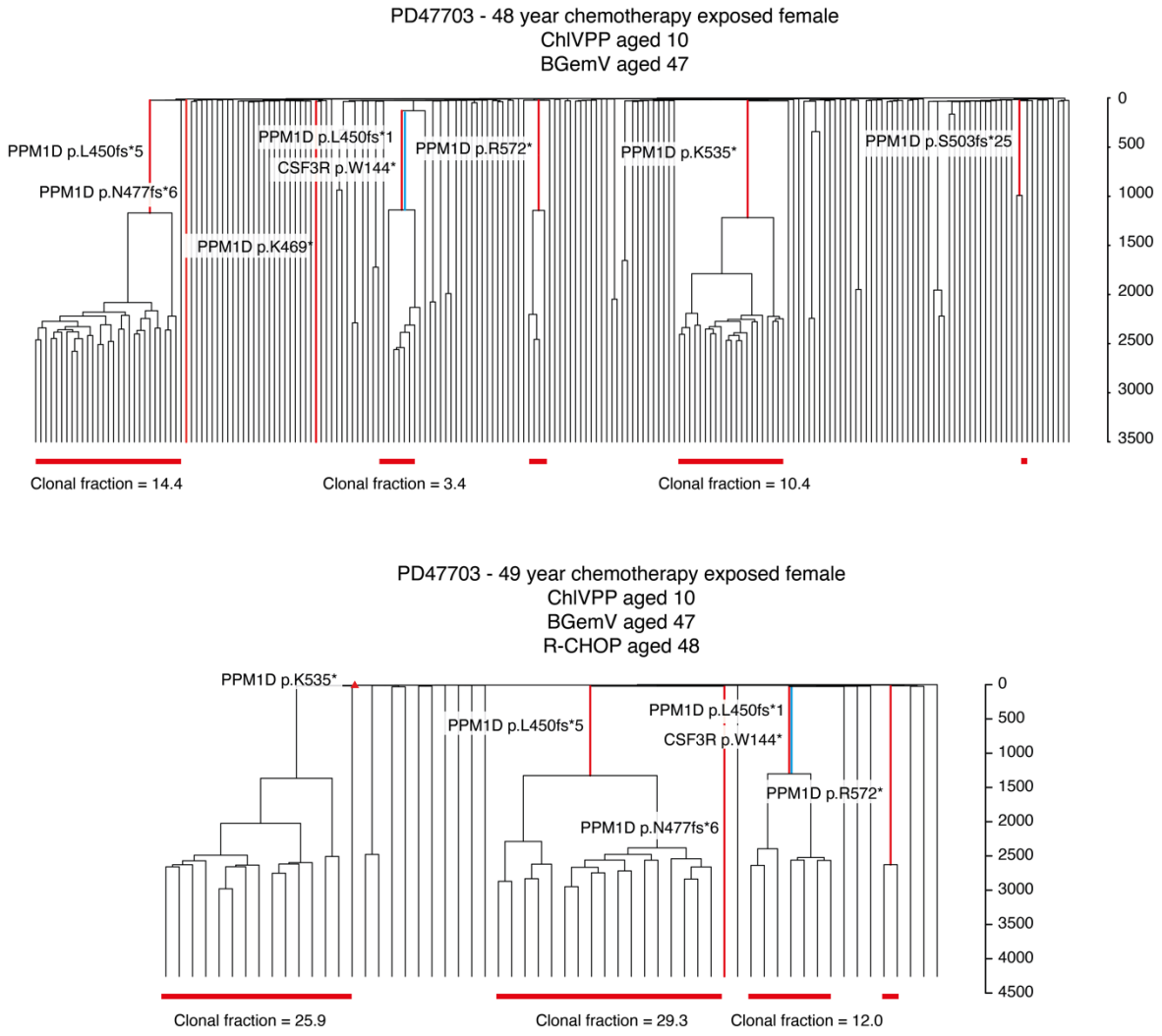


Fig. 5.14| Comparison of PD47703 phylogeny pre and post R-CHOP chemotherapy. Phylogenetic trees reconstructed from somatic mutations for PD47703 timepoint 1 pre-RCHOP aged 48 (top), and PD47703 timepoint 2 post-RCHOP aged 49 (bottom). Axes show mutation number. Branch lengths are adjusted for mean sequencing depth. Probable and definite pathogenic mutations in any one of the top 17 clonal haematopoiesis genes (**Table 3.2**) are depicted, as well as mutations in CHEK2 which have been found to be associated with previous chemotherapy[106] (red). An additional mutation in CSF3R that contributes to an expanded clade is also shown (light blue). Red bars below the phylogeny highlight samples with mutated PPM1D.

Assessment of clonal dynamics in the other phylogenies is less conclusive. Nevertheless, two out of the three smaller phylogenies but none of the matched normal phylogenies capture coalescent events, suggesting a higher degree of oligoclonality in the chemotherapy exposed individuals (**Fig. 5.12**). The phylogeny for PD50307 has one expanded *DNMT3A* clade, which is unusual for a 40 year old. However, this individual has a heavy smoking history (50 pack

years) which may also have contributed to the early onset of clonal haematopoiesis, as well as the *ASXL1* mutation. Smoking is known to increase the risk of clonal haematopoiesis, particularly driven by *ASXL1* mutations[106,205]. It is interesting that PD50308, who like PD47703, received relatively intensive chemotherapy for Hodgkin's lymphoma does not exhibit any clear changes in clonal structure or any chemotherapy associated driver mutations. This is likely partly due to the low resolution of the phylogeny (only 41 samples). In addition, in an analogous process to that seen in normal elderly individuals, the presence of expanding clades carrying driver mutations may only become apparent decades after the clones were initiated. This would explain why PD47703, who received intensive chemotherapy aged 10 and was sampled 38 years later, is the only individual in whom profound clonal changes as a result of chemotherapy are detected.

The other larger phylogeny (PD44579, 173 samples), is certainly more oligoclonal than the 63 year old phylogeny included in the ageing cohort (**Fig. 3.8**). However, the degree of oligoclonality is very similar to the new phylogeny from another haematologically normal 63 year old, albeit with a diagnosis of multiple sclerosis (PD49236). The Shannon Diversity index is 13.9 for PD44579 and 14.8 for PD49236. The phylogeny for PD44579 therefore falls within the normal range of oligoclonality for age. This finding fits with the observation that cyclophosphamide does not cause significant mutagenesis in normal HSPCs, and confers a lower risk for secondary malignancy than other alkylating agents[186]. However, non-mutagenic chemotherapy may still alter the relative balance of fitness effects between clades which we are blind to with our current analysis. One major limitation of the analysis of the chemotherapy-exposed phylogenies other than PD47703 is the lack of baseline assessment of the clonal landscape pre-chemotherapy and future work incorporating this aspect in elderly individuals will be extremely enlightening.

Summary population structure

1. No identifiable bottleneck (or reduction) in HSPC population size as a result of chemotherapy.
2. In young individuals, positive selection of mutations due to chemotherapy may only become detectable years to decades after treatment.

3. Mutagenic chemotherapy causes transient selection on *PPM1D* mutations, and results in fitness effects larger than those observed in the context of age-related clonal haematopoiesis.
4. At least one of the *PPM1D* mutations contributing to clonal expansions in PD47703 in the context of chemotherapy was present in the HSPC population before chemotherapy was given.
5. A *CSF3R* mutation in one of the *PPM1D* expanded clades of PD47703 may have contributed to the prolonged cytopenias experienced by this patient during subsequent RCHOP chemotherapy.

5.5 Summary

In this chapter a survey of the effects of a range of chemotherapeutic agents on normal HSPC mutation burden and population structure has been undertaken. There are several important findings, that warrant further work in this area. The first main finding is that chemotherapeutic agents within the same class have a wide range of mutagenic impacts on normal HSCPCs, from no mutagenic impact at all to conferring thousands of excess mutations. This finding is most notable for the platinum agents and nitrogen mustard derived alkylating agents. The observed range of impacts correlates with myelotoxicity, risk of clonal haematopoiesis and secondary malignancy. However, there is no evidence for reduction in treatment efficacy with the less toxic agents, suggesting they are more specific at targeting malignant rather than normal cells. The other major finding is that there is no evidence for a dramatic reduction in population size as a result of chemotherapy, and that chemotherapy associated oligoclonality occurring in younger individuals is a result of positive selection on DNA-damage response gene mutations, in this dataset *PPM1D* being the most important example. In young individuals it can take years to decades for clonal expansions of driver mutations in these genes to become detectable post chemotherapy. In older individuals in whom relatively expanded DDR gene mutated clones may pre-exist within the HSPC population, the effect of chemotherapy in preferentially expanding these clones may be more readily detectable immediately post treatment.

Chapter 6: Discussion

6.1 Introduction

The work presented in this thesis provides many novel insights into HSC clonal dynamics over the human lifespan. These include first, the observation of universal rapid decrease in clonal diversity after the age of 70. Second, that three quarters of clonal expansions in elderly individuals have no known driver mutations. And third, that clonal expansions start decades prior to becoming detectable, typically originating in childhood or early adulthood. In addition, other age-related changes in human HSCs have been characterised at single cell resolution, including mutation accumulation and telomere attrition. Finally, studying normal HSPCs in individuals previously exposed to chemotherapy has revealed a wide variation of mutagenic impact between agents within the same class as well as insights into the impact of chemotherapy on HSC clonal dynamics. The findings presented in this thesis will be discussed with regard to how they have aided our understanding of the relevant fields of research, as outlined below.

Key points to be covered in this chapter

1. Novelty of methodological approach.
2. Overview of simple model presented to explain HSC clonal dynamics.
3. Implications of work presented for understanding HSC ageing.
4. Implications of work presented for understanding development of haematological malignancy.
5. Implications of work presented for understanding impact of chemotherapy on normal HSCs.
6. Future work.

6.2 Novelty of methodological approach

Single cell resolution at stem cell level

The main difference between this work and the majority of clonal haematopoiesis literature to date is the sequencing approach used. This study sequenced single HSC/MPP-derived colonies, whereas the vast majority of studies on clonal haematopoiesis have sequenced DNA from bulk blood cell populations[22,79,81,206,207]. Sequencing of bulk blood populations has the advantage that samples from many more individuals can be assessed. However, most bulk sequencing studies are only capable of detecting mutations in clones that contribute more than around 5% blood production (VAF >2%), so is only able to identify the one or two most dominant clones. In contrast, the approach used here provides resolution at the level of individual HSC/MPPs. This allows us to obtain data from many 'singleton' colonies, whose clonal fraction is considerably lower than what could be detected with bulk sequencing. In addition, this study provides the first assessment of clonal dynamics in the stem cell compartment, eliminating any biases in detected clone size due to selection during differentiation.

Genome wide coverage across many samples

Most studies of clonal haematopoiesis have performed targeted sequencing of the exome or fifty to a few hundred known cancer genes, with analyses focussed on known drivers of myeloid leukaemia (approximately 100 genes). This means evidence for selection acting outside these genes has not been previously well assessed. In contrast whole genome sequencing over 3,500 colonies, as performed here, has provided statistical power to identify positive selection across the genome in an unbiased way. It has also allowed the identification of clonal expansions that occur in the absence of known drivers.

Accurate phylogeny reconstruction

Studies using bulk sequencing have not been able to determine whether driver mutations identified are from the same clone or not, or estimate their time of acquisition. Whole-genome sequencing of single cell colonies allows reconstruction of highly accurate phylogenies. The phylogeny approach provides precise information on mutation co-occurrence, and allows inference of the time of onset of clonal expansions to be inferred. The

wealth of information on clonal dynamics contained within the large phylogenies from elderly individuals has allowed the development of a simple and insightful model for HSC clonal dynamics across the human lifespan.

6.3 A simple model to explain HSC clonal dynamics

Overview of the Approximate Bayesian Computation approach

An Approximate Bayesian Computation (ABC) approach[131,208] was used to determine the values of key model parameters based on the pattern of coalescent events (or clonal relationships) in the elderly phylogenies. To summarise the approach, in the first phase 100,000 different simulations of long-term HSC populations are performed per individual. Each simulation followed the same assumptions: 1) Constant HSC population size over life of 100,000 (derived from young adult phylogenies and consistent with other estimates in the literature[15,143]); 2) HSC generation time of 1 year (derived from telomere loss data and consistent with other estimates in the literature); 3) Constant entry of driver mutations into the HSC compartment over life; 4) Fitness effects of driver mutations drawn from a gamma distribution; 5) Driver fitness effects constant over time.

The values for some key parameters in the model were not known, namely the rate and shape of the gamma distribution of fitness effects and the rate of driver mutation entry into the population. In order to provide estimates of these from the observed phylogenies, each simulation took a random draw from flat uninformative prior distributions for each parameter. Informative summary statistics calculated from phylogenetic trees produced from the simulations could then be compared with the summary statistics from the real observed phylogeny for each individual in turn. From the top 1% of simulated phylogenies that best match the observed data, the posterior distributions of the parameters of interest can be extracted. In all cases the posterior distributions were a well-defined subspace of the prior distribution, showing the observed, or real, phylogenetic trees contain considerable information about the key parameters of interest. Estimation of the key parameters in this way allows an 'optimal' model for HSC clonal dynamics across life to be determined. The model developed is one of the major achievements of this thesis.

Key features of the derived model

The ABC approach described above allowed the first estimation of the distribution of driver mutation fitness effects in the HSC compartment of normal individuals. It found there are many drivers with moderate fitness effects (5-10% additional growth per year) and a long tail of rare drivers with higher fitness effects (> 10% additional growth per year). It also allowed the rate of driver mutation acquisition in the HSC population to be estimated at 2.0×10^{-3}/HSC/year. Together with the information on HSC population size and generation time obtained from the young adult phylogenies, this work therefore provides estimates of all the important parameters needed to optimally model clonal dynamics in the human LT-HSC compartment (**Fig. 6.1**).

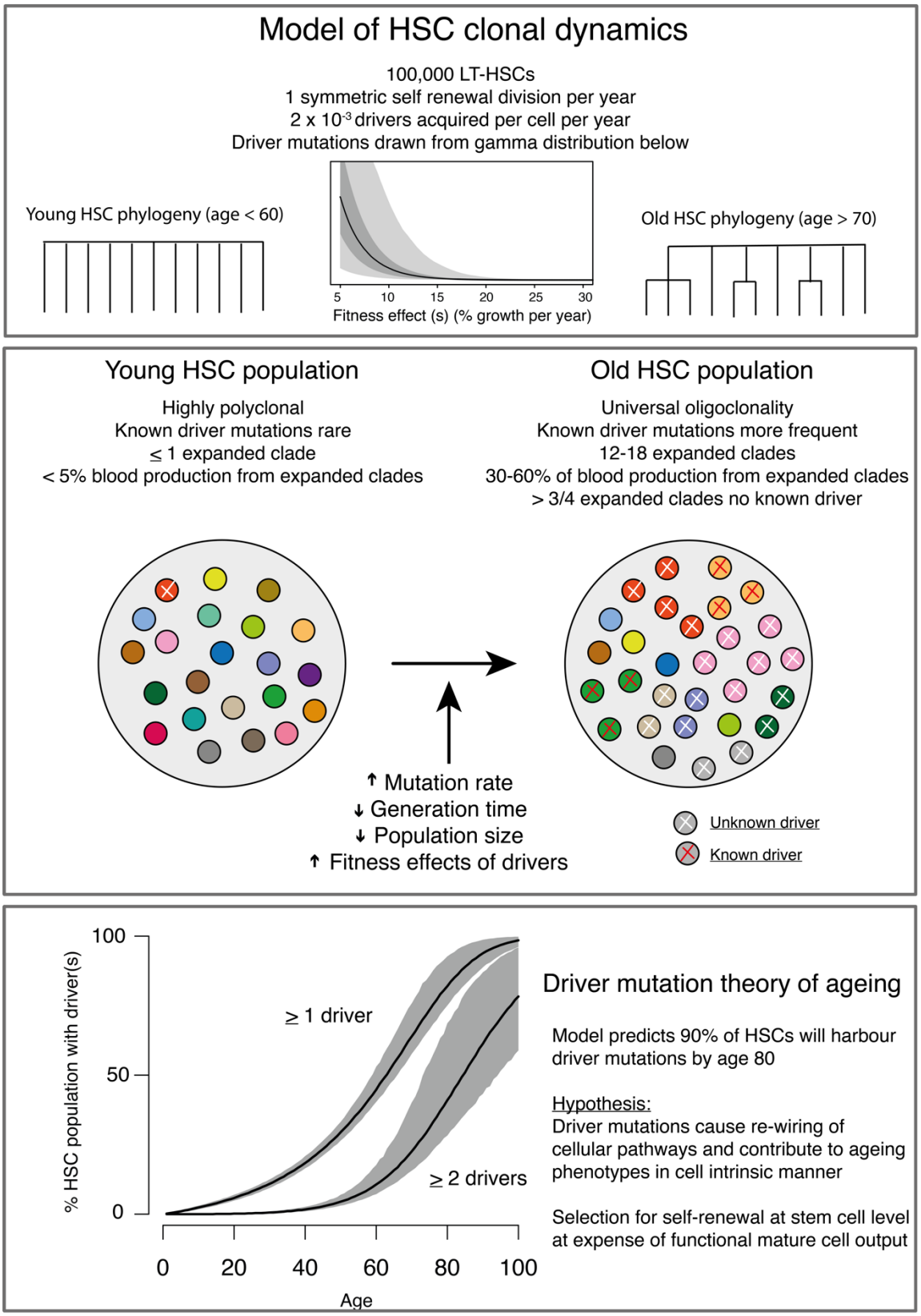


Fig. 6.1| Overview of model of HSC clonal dynamics. Top panel shows key parameters of model of HSC clonal dynamics and the most credible gamma distribution of driver fitness effects. Middle panel outlines the main findings from the young and old phylogeny reconstruction. The four factors below the main arrow are those that would be predicted to speed up the transition to oligoclonality

with age. Bottom panel gives overview of driver mutation theory of ageing with plot showing median and 95% posterior intervals for the percentage of HSCs with drivers calculated yearly for 10,000 HSC population simulations run utilising the optimal parameter values for driver acquisition rate and fitness effects derived from the ABC modelling approach.

Reassuringly the narrow window of optimal parameter estimates (defined using only the elderly phylogenies), generates simulations that match the inflection point of markedly reduced clonal diversity over the age of 70. The optimal model also facilitates exploration of other difficult to determine parameters, allowing for example the prediction that by age 80, 90% of HSCs will harbour at least 1 driver mutation and 50% will harbour 2 or more. Of course, native human haematopoiesis is more complex than this simple model, with its various compartments of stem and progenitor cells, lineage biases, epigenetic change and interconnected microenvironment. Despite these limitations, the model provides a useful framework for understanding HSC clonal dynamics across the human lifespan.

6.4 HSC ageing

'Driver mutation' theory of ageing

One fundamental conundrum of ageing is how the gradual accumulation of cellular 'wear and tear' can lead to a sudden decline in organ and organismal function near the end of lifespan (over the age of 70 in humans)[46]. The model presented above provides insights into how lifelong, constant accumulation of molecular damage (in the form of somatic mutations), can lead to abrupt deterioration in the haematopoietic system after the age of 70 years, potentially providing one example of how non-linear age-related effects can occur. Second the model can explain how ageing can be universal but variable between individuals, as the wide pool of possible driver mutations means that although it is inevitable that some clones will dominate the HSC population by old age, the properties of the specific drivers within these clones can determine inter-individual variation in blood counts and ability to tolerate chemotherapy for example. We therefore propose a 'driver mutation' theory of ageing, where the term 'driver' is used to describe any somatic mutation under positive selection in the HSC population. This theory suggests that at least some age-related loss of function within the haematopoietic system with age, is due to the accumulation of these positively selected

'driver mutations'. Although these mutations confer enhanced fitness (in the form of self-renewal) at the stem cell level, the impact on the tightly orchestrated production of functional mature blood cells can be negative.

Intriguingly this novel 'driver mutation' theory of ageing was also supported by data from one individual in the 'chemotherapy-exposed' cohort who had a large clone harbouring a loss of function mutations in both *PPM1D* and *CSF3R*. Loss of function *CSF3R* mutations are commonly seen in individuals with congenital neutropenia and therefore the presence of this clone could in part explain the prolonged cytopenias she developed with subsequent chemotherapy. This provides a potential proof of principle for the idea that mutations under positive selection (either in their own right or as passengers) can contribute to loss of function within the haematopoietic system as a whole.

The 'driver mutation' theory of ageing also fits with the finding that age-related change in HSCs is largely cell intrinsic and that there is heterogeneity of aged phenotypes within the population. For example, the fact that a small proportion of HSCs in elderly mice retain a youthful function[65,66,157] could be explained by the fact that these represent the small proportion that have not acquired a driver mutation by old age. That age-related phenotypes might be at least in part determined by somatically acquired mutations would also fit with the observation that both transplantation into young mice[65,69,157] and systemic rejuvenation strategies are unable to restore function in aged HSCs[209].

Impact of microenvironmental changes with age

Despite the simple model above being sufficient to explain the observed clonal dynamics in elderly individuals, the impact of microenvironment changes with age cannot be completely discounted. It is widely accepted that ageing results in changes in the bone marrow niche. For example, ageing is associated with changing cytokine concentrations in the bone marrow. However, the elderly phylogenies presented report on *relative fitness* of different clones within the same individual; namely, HSCs competing with one another in the same haematopoietic microenvironment. For microenvironmental ageing to promote the loss of clonal diversity we observe, it would still have to act in a clone-specific manner. Interestingly, our data show that all of the expanded clades began their expansion in the first half of life –

this suggests that their selective advantage is lifelong, emerging before the local marrow microenvironment has itself undergone age-related remodelling.

Nevertheless increased signalling of specific cytokines could synergise with specific driver mutations (as has been shown experimentally for *TET2* and IL6[210] as well as *DNMT3A* and IFNγ[211]), conferring an increased proliferative advantage on these clades, or cause a more general increase in cell division rates across the whole population. Other microenvironmental changes include the accumulation of adipocytes as well as remodelling of structural components such as the vasculature, extracellular matrix and bone. These are likely to alter the niche for HSCs and potentially shape the selective landscape the clones are competing in, which may alter clonal dynamics. As a further complication, it is entirely feasible that clones with driver mutations actively remodel their own niche – this has been documented in mouse models of *Tet2* mutations, for example[212].

Functional genomics implications

The observation that clonal diversity diminishes sharply after the age of 70 is interesting from a functional genomics perspective. One explanation for this is that our historic demographic structure has exerted evolutionary pressure on the germline genome to find strategies to reduce blood cancer risk up till the ages of 50-65 years. However, the relatively low historic contribution of individuals over this age to successful reproductive output substantially reduces that evolutionary pressure in older individuals.

With this evolutionary perspective, the model above allows us to define the parameters that determine the age at which clonal diversity declines: 1) The population size of HSCs and generation time (average time between symmetric self-renewals); 2) Rate of acquisition of driver mutations into the HSC compartment (a function of both the overall mutation rate and the fraction of those mutations that confer selective benefit); 3) The size of the fitness benefit conferred by driver mutations. One prediction from this work is that each of these parameters has had sustained (germline) evolutionary pressure to maintain a high degree of clonal diversity until the age of 60-70 years.

Ageing across species

Supportive evidence for the role of evolution in determining the age of onset of oligoclonality comes from considering other mammalian species with very different lifespans. For example, somatic mutation rates have evolved such that the end-of-life mutation burden is remarkably similar across different mammals, despite there being 50-fold variation in lifespan[213]. Considering HSC-specific parameters, shorter lived species such as mice have a smaller HSC population (~5,000-10,000)[214,215], with a shorter HSC generation time (~5 weeks)[155,156,216] – smaller population size and more rapid generations both lead to accelerated opportunity for clonal expansions, but presumably a maximal lifespan of 2-3 years in a mouse limits the risk that these will result in decreased reproductive output. Some degree of decreased clonality has already been observed in elderly mice[217] and macaques[218], providing further support for the 'driver mutation' theory of ageing.

Ageing in other tissues

The extent to which the onset of oligoclonality may be contributing to age-related changes in other tissues is another interesting discussion point. Solid tissues present more of a challenge for the analysis of organ wide clonal diversity than blood. The haematopoietic system is unique in being well-mixed – when the variant allele fraction of mutations in a single bone marrow draw is compared with that of peripheral blood, there is a strong correlation[15].

In contrast, stem cell clones in solid organs show significant spatial organisation[219]. It is therefore much more difficult to obtain a truly random sample of stem cells from solid organs. However, studies using non-random sampling of clonal stem cell units in the liver[220] and colon[55] have shown high levels of polyclonality and stem cell diversity in non-diseased young individuals. This work has not been extended yet to many elderly individuals, but studies in the diseased setting (inflammatory bowel disease[221] and cirrhosis[220]) have shown larger clone sizes (millimetres to centimetres in size) and reduced diversity of stem cells, driven by a combination of selection for protective driver mutations, and the need to regenerate across damaged regions.

In many normal tissues studied to date there is also evidence of convergent evolution with the same genes recurrently mutated and positively selected in independent clones. This can result in 20-80% of all epithelial cells in skin[222,223], oesophagus[224,225], endometrium[115] or bronchus[54] carrying mutations in specific driver genes in the elderly, similar to the proportions of HSCs found in expanded clades in elderly blood. One potential significant difference between driver mutation selection in blood compared to solid organs however is the number of drivers that are under sufficient selection to expand within the population. A recent study by Poon *et al*[183]*,* which analysed patterns of synonymous mutations in blood and oesophagus bulk targeted data, also found evidence that many as yet unknown gene mutations are driving clonal expansions in blood. However, this was not the case in oesophagus where their work predicted the vast majority of driver mutations lie within just two genes (*NOTCH1* and *TP53*).

6.5 Development of haematological malignancy

Role for novel driver mutations

The observation that there are many as yet unknown drivers contributing to clonal expansions in blood potentially has important implications for our understanding of the development of haematological malignancy. Previous studies of clonal haematopoiesis, which have predominantly focussed on known cancer genes, have consistently shown an increased risk of subsequent blood cancer in individuals carrying known driver mutations. However, the risk arising from expanded clones without known driver mutations has not been comprehensively studied. An analysis of 11,262 bulk whole genome sequences from the deCODE cohort provides the most informative data to date[82]. The study identified three groups of subjects: (1) those without detectable clonal haematopoiesis (acknowledging that bulk WGS could only identify clones with VAFs larger than 10-20%); (2) those with clonal haematopoiesis carrying known drivers; and (3) those with evidence for clonal haematopoiesis but no known drivers. Interestingly, the latter two groups both had elevated risk of all-cause mortality and future blood cancer. The elevated risk was similar whether or not known drivers were present[82].

While these data do suggest that clones without known driver mutations can contribute to leukaemic transformation, it is also possible that their presence denotes a selective pressure

or microenvironmental effect that is conducive to clonal outgrowth, with future malignancies arising from an independent clone not detectable at the time of WGS. At present there is no definitive evidence that the clones without known driver mutations have themselves an increased risk of malignant transformation. For example, screening 534 published AML genomes[146,147,226] for variants in the two relatively new genes described in this work, identified only one possible oncogenic mutation in *ZNF318* and no mutations in *HIST2H3D*.

Although the data presented in this thesis and in another companion study[133] do not provide definitive evidence that unknown driver events contribute to myeloid malignancies, there are suggestions that some known drivers are more typically acquired as a 'second hit', after an initial clonal expansion earlier in life. This is potentially particularly true for the phylogenies containing spliceosome mutations in Fabre *et al*[133], where the three of the four identified spliceosome mutations occurred as second events, with LOY being the probable first event in one case. This could explain why spliceosome mutations are more common in men (if LOY is a common first hit), and almost exclusively only detectable in older individuals.

One other possibility is that some clones with unknown driver mutations may be protective against the development of haematological malignancy. Unknown driver clones with high fitness effects could outcompete clones containing known driver mutations later in life. This may explain why there is a reduction in risk of myeloid malignancy in the very elderly (age >100). Very elderly individuals likely have only one or two dominant clones, as shown in one 115-year old[85] and our modelling work, that have / would outcompete nascent malignant clones. Another interesting clinical observation in this regard is that acute leukaemias in the very elderly can be relatively slowly progressive, with individuals able to survive for 1 year or more with supportive care only. This is not the case in young individuals in whom acute leukaemias typically rapidly progress. This may reflect competition from other non-malignant clones with high fitness effects. A number of other possible explanations exist, for example that acute leukaemias in the elderly are more likely to occur as a result of progression from a myelodysplastic syndrome.

Origins of malignancy early in life

The other key finding of the ageing work that has implications for our understanding of the development of haematological malignancies is that the clonal expansions we identified in elderly blood were universally acquired decades prior to sampling, typically in childhood or young adulthood, although none were acquired during early embryonic or foetal life. In contrast another parallel study focussing on the clonal dynamics in patients with myeloproliferative neoplasms[128] found a high proportion of the mutations driving the malignant clone had been acquired *in utero*, particularly in those individuals in whom disease was diagnosed at a young age. The distribution of fitness effects we observed in normal individuals results in clonal expansions that take decades to become detectable. This suggests that for myeloid malignancies to be acquired at a young age requires either 1) The early acquisition of a highly potent but very rare driver (for example *BCR-ABL1* or *PML-RARA* translocations); 2) The acquisition of the first hit *in utero* (for example *JAK2* mutations in the myeloproliferative neoplasms[128]); or 3) The rare event of multiple hits being acquired in rapid succession in a single clone (observed in our optimal model in approximately 1 in 10,000 simulations). This might also explain why acute leukaemias acquired in young individuals are more rapidly progressive, as they potentially require higher fitness drivers to come to dominate in the population at a younger age.

6.6 Impact of chemotherapy on normal HSCs

Variable mutagenic impact of agents within the same chemotherapeutic class

One of the most interesting findings from the chemotherapy-exposed cohort was the huge variation in mutagenic impact of agents within the same class (**Fig. 6.2**). This variation in mutagenicity was true for both the platinum agents and nitrogen mustard derived alkylating agents. In addition, it is highly likely that the same is true for the monofunctional alkylating agents, with procarbazine being highly mutagenic, but dacarbazine not. There is a tentative mechanistic proposal for why cyclophosphamide might be expected to have a reduced mutagenic toxicity on stem and progenitor cells, as it is thought to be inactivated by high levels of ALDH1 specifically in these cells[227]. However, the pharmacological mechanisms underlying the variation in platinum agent mutagenicity and potentially also the variation in monofunctional alkylating agent toxicity are not at all well understood.

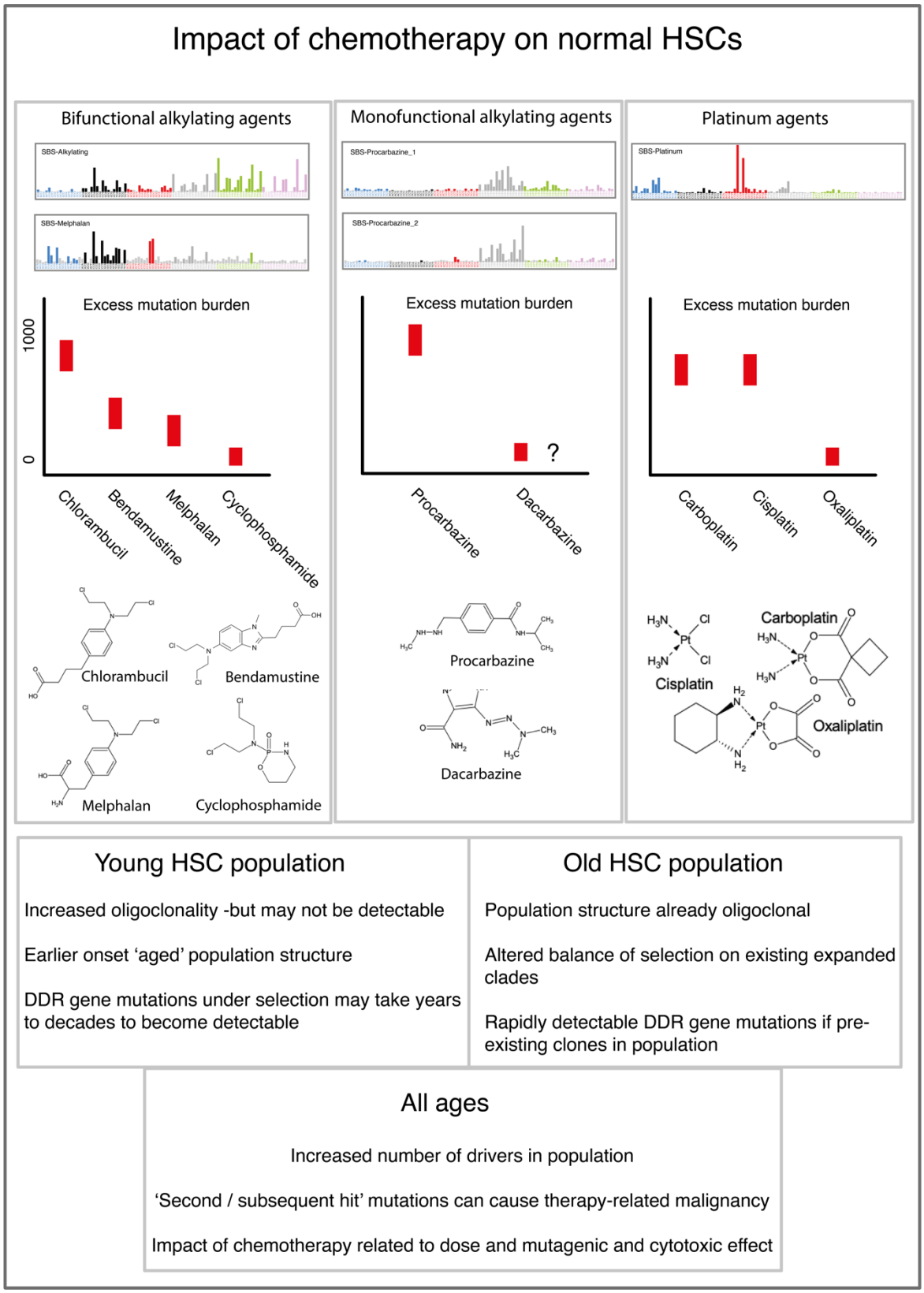


Fig. 6.2| Overview of impact of chemotherapy on HSCs. Top panels show mutational signatures, approximated excess mutation burden and molecular structure of chemotherapeutic agents in the three main classes of drug where one or more agents was found to be mutagenic to normal HSCs.

Of not the excess mutation burden is likely to be dependent on cumulative dose. Definitive data is awaited regarding the mutagenic impact of Dacarbazine but evidence from toxicity and impact on germline suggests is confers very few excess mutations. It is unclear from our data whether cisplatin or carboplatin is more mutagenic to normal HSCs. Bottom panel highlights findings or predictions regarding the impact of chemotherapy on young and old HSC populations.

The finding of variable mutagenic impact between chemotherapeutic agents in the same class has important clinical implications. There is no evidence for reduced efficacy of oxaliplatin as compared to cisplatin and carboplatin in paediatric germ cell tumours for example, where oxaliplatin is used second line in individuals that have relapsed post carboplatin / cisplatin. In addition, a recent (unpublished) clinical trial of switching procarbazine for dacarbazine in EscBEACOPP has found reduced blood transfusion requirements and less impact on fertility in the context of non-inferior clinical outcomes. These observations suggest there is room to improve current treatment regimens by substituting mutagenic agents for non-mutagenic agents in the same class, so reducing the risk of secondary leukaemia and other adverse effects of excess mutation acquisition. Additional clinical trials of switching out mutagenic agents are clearly warranted, especially in the paediatric setting where the risk of secondary MDS/AML post treatment for neuroblastoma is 5-10%[228].

Impact of chemotherapy on clonal dynamics and ageing

The simple model of HSC clonal dynamics also allows us to predict how chemotherapy may impact HSC population structures (**Fig. 6.2**). Firstly, chemotherapy has been shown to result in increased HSC turnover and reduced generation time in mice treated with 5FU[104,105]. This would be predicted to speed up the growth of pre-existing expanded clones, so reducing the age at which oligoclonality becomes detectable and 'ageing' the organ system. In addition, mutagenic chemotherapy would be predicted to both increase the rate of driver acquisition in the HSC population and dramatically alter the selective landscape, favouring driver mutations that allow escape from DNA-damage-response pathways, as seen in PD47703[106–108].

The phylogenies created from individuals post-chemotherapy fit with observations made in the ageing cohort that clones under positive selection can take decades to expand to a detectable level. In young individuals exposed to chemotherapy the impact of the altered

selective landscape can take years to decades to become detectable. *PPM1D* mutations were found to have a fitness effect of 100% over 1 year during which 5 cycles of RCHOP were given. This is much higher than the fitness effects observed in the fittest clones of normal individuals, but not high enough to cause detectable clones in the first few years post chemotherapy in young individuals, whose baseline haematopoiesis is highly polyclonal. Although it is likely that more intensive and mutagenic regimens cause a more significant selective pressure than RCHOP, there was no evidence of selection of DNA-damage-response gene mutations in the small phylogeny for PD50308 (29 year female) 2 years post the intensive and mutagenic regimen escBEACOPP. Nevertheless, clonal haematopoiesis with known drivers has been identified in 2-4% of patients treated for paediatric cancer, a much higher level than is detectable in untreated children[106,108], supporting the view that chemotherapy brings forward the age at which oligoclonality is detectable.

In elderly individuals, it is difficult to assess the impact of chemotherapy on clonal dynamics without having a 'pre-chemotherapy' phylogeny for comparison. One would predict that chemotherapy might alter the relative balance of selective pressures on existing clones, so altering the clonal landscape. This would be particularly likely in the presence of pre-existing clones with mutations in DNA-damage-response genes, which would be predicted to preferentially expand. These predictions are supported by work showing that in older individuals exposed to chemotherapy, the expanded driver mutation clones detectable post-chemotherapy were almost always detectable pre-chemotherapy[106]. In addition, whereas some clones expanded as a result of chemotherapy others declined in size, in keeping with a change in the relative balance of selective pressures.

Risk of secondary myeloid malignancy

One interesting feature of secondary myeloid malignancies (therapy-related MDS and AML), is that there is typically a 'time-window' in the 5-10 years following treatment where the risk is elevated, with age over 50 being associated with a higher risk[109]. Treatment for Hodgkin's disease is associated with a 20—40-fold increased risk of AML, while patients treated for Non-Hodgkin's lymphoma (NHL) have a lower, 2-15-fold increased risk. The risk drops to background population rates 10 years post treatment. The phylogeny data over life, shows that all individuals acquire mutations in driver genes at a young age and that clones with single

drivers typically take a lifetime to expand to detectable levels. Therefore, chemotherapy most likely contributes to the development of secondary myeloid malignancies by causing either multiple leukaemogenic hits in the same clone, or by contributing second or subsequent hits to clones that already possess one or more drivers. These clones with multiple drivers derived from chemotherapy would be predicted to have high fitness effects and expand within the timeframe of 5-10 years. So, if there is no malignant clonal outgrowth in this time window then chemotherapy has not resulted in a multiply mutated clone. It also explains why the risk is higher in individuals over the age of 50, as they will harbour many more cells with single drivers than younger individuals, providing a greater opportunity for chemotherapy to cause a second / subsequent hit.

6.7 Future work

There are a number of exciting avenues of future work that is being or will be pursued as outlined below:

1. Analysis of scRNA sequencing data from a subset of colonies in the ageing cohort, with the goal to investigate age-related and clade specific changes in the transcriptome.
2. Analysis of methylation data from a subset of colonies in the ageing cohort, with one goal being to identify whether there are unifying epigenetic changes across clonally expanded clades.
3. Cross-species assessment of HSC clonal dynamics, aiming to include (amongst others) mouse, naked mole rat and elephant; with the goal to determine if oligoclonality is found in the HSC populations of all mammals towards the end of healthy lifespan.
4. Analysis of structural variants and copy number variants in the chemotherapy cohort to determine if any specific agents are associated with elevated burdens of these mutation types.
5. Analysis of the impact of chemotherapeutic agents on mature blood cells (B cells, T cell, monocytes and granulocytes from our chemotherapy cohort), with the goal to determine if these mature cells types are more or less impacted by specific chemotherapeutic agents as compared to HSPCs.
6. Characterising the mutagenic impact of the range of regimens used to treat Hodgkin's lymphoma in a new cohort of 12 patients (EscBEACOPP, EscBEACOP-Dac, ABVD), with

the goal to test the prediction made here that dacarbazine is less mutagenic than procarbazine.

7. Characterising the mutagenic impact of the range of platinum containing regimens used to treat paediatric solid cancers, with the goal to determine if oxaliplatin could replace cisplatin and / or carboplatin in current regimens.
8. Assessment of the mutagenic impact of a wider range of chemotherapeutic agents, with the goal to provide a compendium of drug impacts on normal HSPCs.

6.8 Conclusion

In conclusion, the work presented in this thesis has generated a number of novel insights into HSC clonal dynamics over the normal lifespan. These insights have important implications for both our understanding of ageing and the development of haematological malignancies. The derived model of HSC clonal dynamics provides a useful framework for understanding the impact of both ageing and perturbations to the haematopoietic system. The work on the impact of chemotherapy has highlighted clinically relevant findings that are already leading to important further work to improve the outcome of patients treated for both haematological and solid cancers.

Appendix 1: Supplementary Simulations

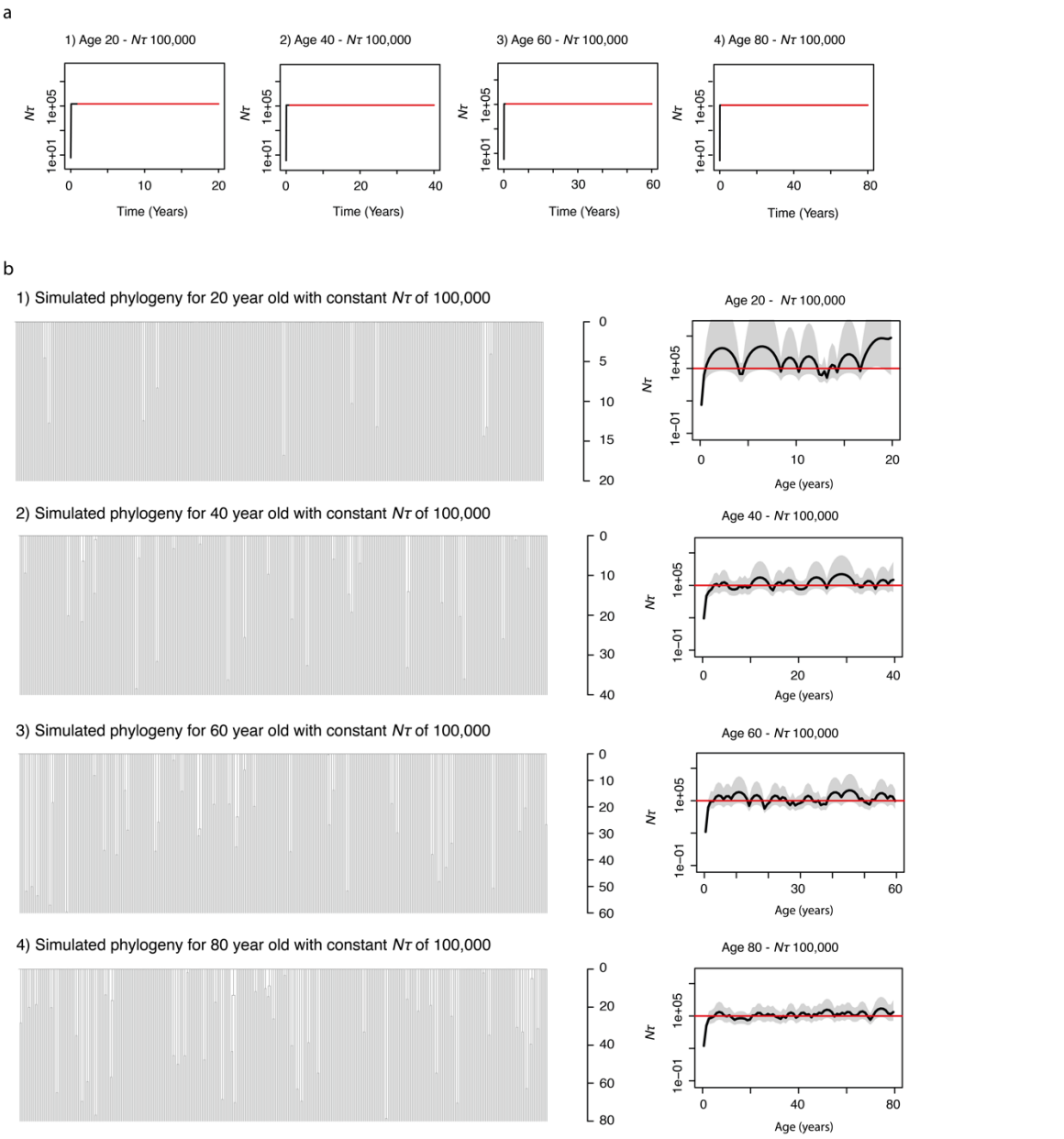
In all the simulated phylogenies illustrated below, the R package *rsimpop* was used to simulate a full neutrally evolving HSC population of size *N*. At a given age 380 cells were sampled at random from the full population to allow creation of comparable phylogenies to those we have obtained from real HSC/MPPs. In all simulations the generation time (τ) was set at 1 year meaning $N\tau = N$. A *phylodyn* plot is also shown for each phylogeny to show how accurately the population trajectory could be recreated from the pattern of coalescent events. In *phylodyn* plots the downward dips in the trajectory (black line) represent coalescent events in the phylogeny.

Effect of population size

Increasing *N* reduces the number of coalescent events per unit time in the phylogeny of cells with a fixed generation time sampled from an individual of a given age (**Fig. 1.4**). At age 30 there is loss of resolution in the *phylodyn* output between $N\tau = 500{,}000$ and $N\tau = 750{,}000$.

Effect of age

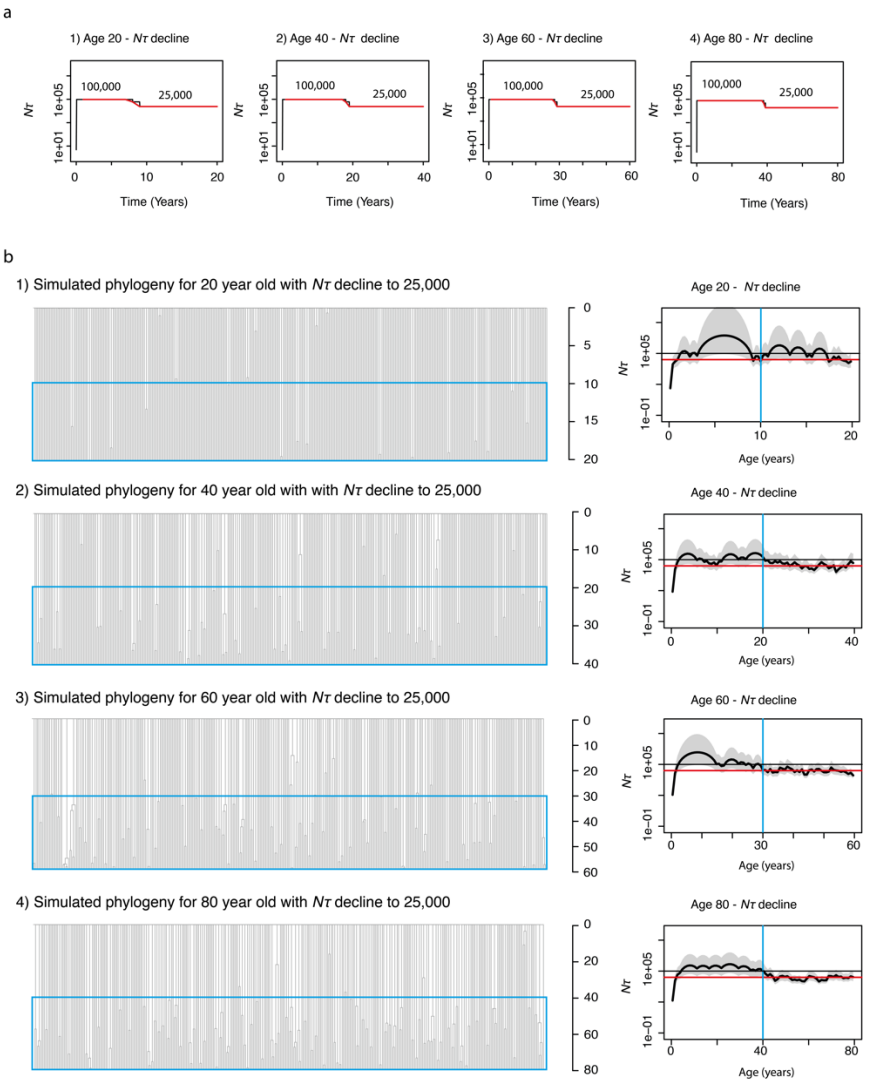
Increasing age allows a more accurate estimate of $N\tau$ due to the higher number of coalescent events per unit time (**Appendix Fig. 1**). This increase in the number of coalescent events per unit time for a given population size at homeostasis occurs as a result of genetic drift.



Appendix Fig. 1|Effect of age. a, Trajectories of $N\tau$ used as input to *rsimpop* for the simulations to create phylogenies in b. Note the Y axis depicting $N\tau$ is on a log scale. **b,** Phylogenies created by randomly sampling 380 cells from the final full simulated population of 100,000 cells at between age 20 (Phylogeny 1), age 40 (Phylogeny 2), age 60 (Phylogeny3) and age 80 (Phylogeny 4). Each simulation has a constant $N\tau$ of 100,000 In all cases $N\tau$ is the same as the population size (N), as the generation time (τ) is 1 year. The *phylodyn* trajectories to the right of each simulated phylogeny use the pattern of coalescent events to recover the input trajectories for $N\tau$.

Population decline

A decline in population size is reflected by an increase in the number of coalescent events captured per unit time as compared to when the population was larger (**Appendix Fig. 2**). Again, the older the individual the more accurately *phylodyn* is able to recover the true simulated population size trajectory. Decreases in $N\tau$ to less than 25,000 can be reasonably accurately captured by *phylodyn*. In younger individuals this is best observed as an increase in the frequency of bumps in the trajectory.

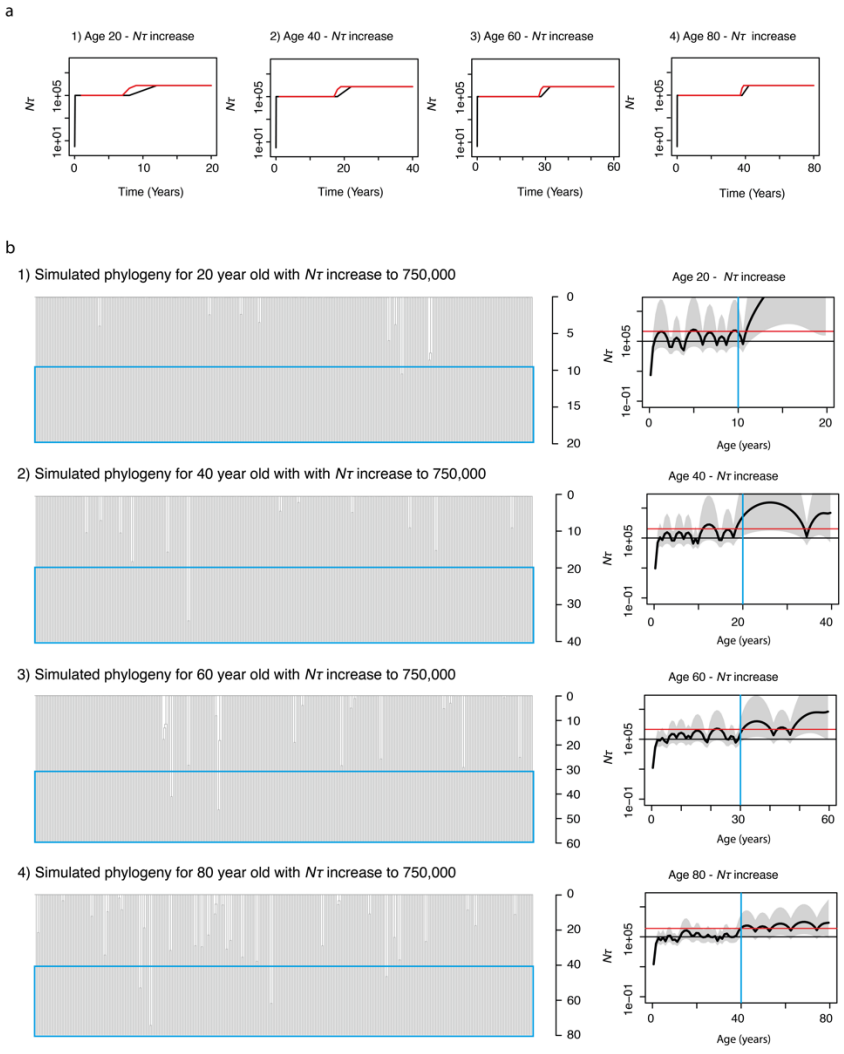


Appendix Fig. 2|Effect of population decline. a, Trajectories of $N\tau$ used as input to *rsimpop* for the simulations to create phylogenies in b. Note the Y axis depicting $N\tau$ is on a log scale. **b,** Phylogenies created by randomly sampling 380 cells from the final full simulated population of 25,000. Each simulation has an initial $N\tau$ of 100,000 with a decline to 25,000. In all cases $N\tau$ is the same as the population size (N), as the generation time (τ) is 1 year. The blue boxes indicate the period of time

in which the population size is decreased. The *phylodyn* trajectories to the right of each simulated phylogeny use the pattern of coalescent events to recover the input trajectories for $N\tau$. The blue line marks the time of change in $N\tau$.

Population growth

An increase in population size is reflected by a decrease in the number of coalescent events captured per unit time as compared to when the population was smaller (**Appendix Fig. 3**). Again, the older the individual, the more accurately *phylodyn* is able to recover the true simulated population size trajectory. Increases in population size to over 500,000 result in a loss of resolution (and overestimation of $N\tau$ in individuals < 40). In younger individuals the change in population size is best observed as a reduction in the frequency of coalescent events (bumps in the trajectory), but the magnitude of the change cannot be accurately determined.

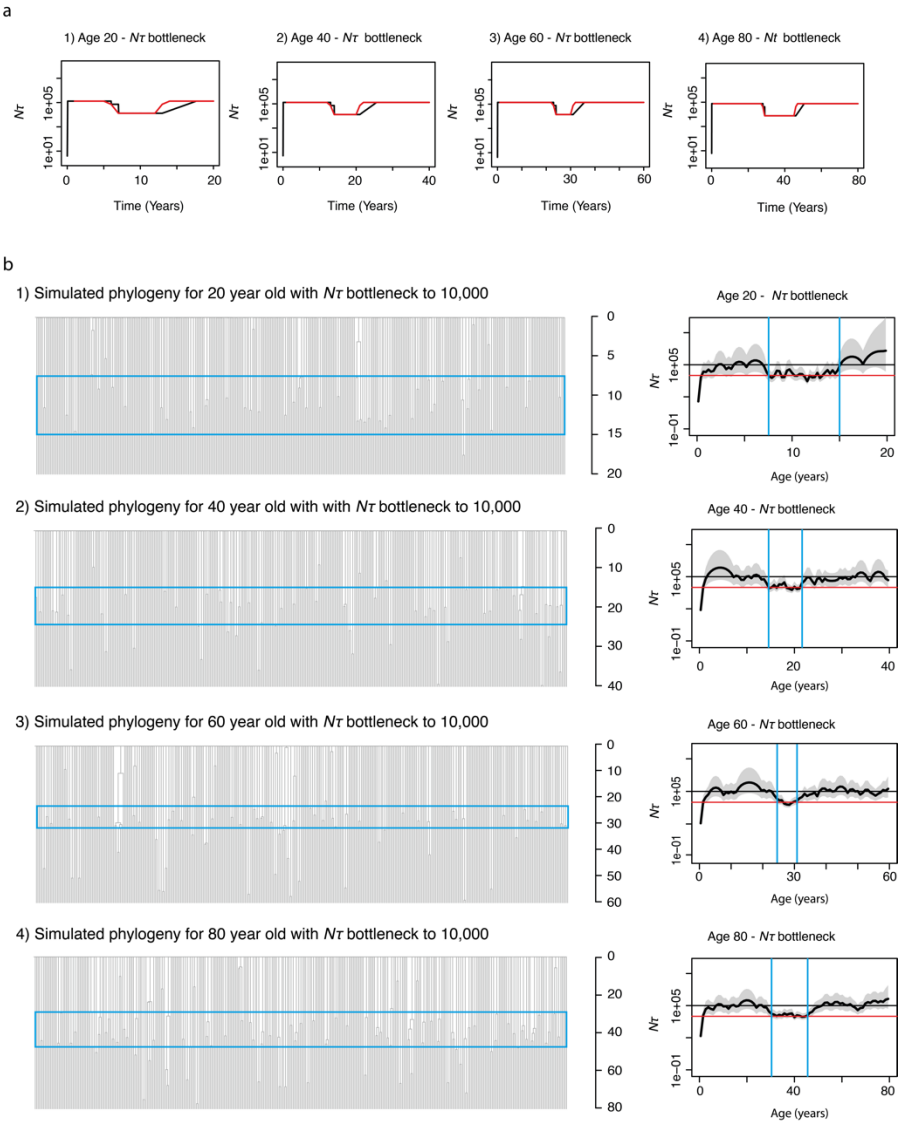


Appendix Fig. 3|Effect of population increase. a, Trajectories of $N\tau$ used as input to *rsimpop* for the simulations to create phylogenies in b. Note the Y axis depicting $N\tau$ is on a log scale. **b,**

Phylogenies created by randomly sampling 380 cells from the final full simulated population of 750,000. Each simulation has an initial $N\tau$ of 100,000 with an increase to 750,000 in midlife. In all cases $N\tau$ is the same as the population size (N), as the generation time (τ) is 1 year. The blue boxes indicate the period of time in which the population size is increased. The *phylodyn* trajectories to the right of each simulated phylogeny use the pattern of coalescent events to recover the input trajectories for $N\tau$. The blue line marks the time of change in $N\tau$.

Population bottlenecks

'Bottlenecks' in the population represent periods of time with a reduced population size compared to baseline. These can be recovered accurately by *phylodyn* at all ages, given a reduction to in $N\tau$ from 100,000 to 10,000 during the bottleneck period (**Appendix Fig. 4**).

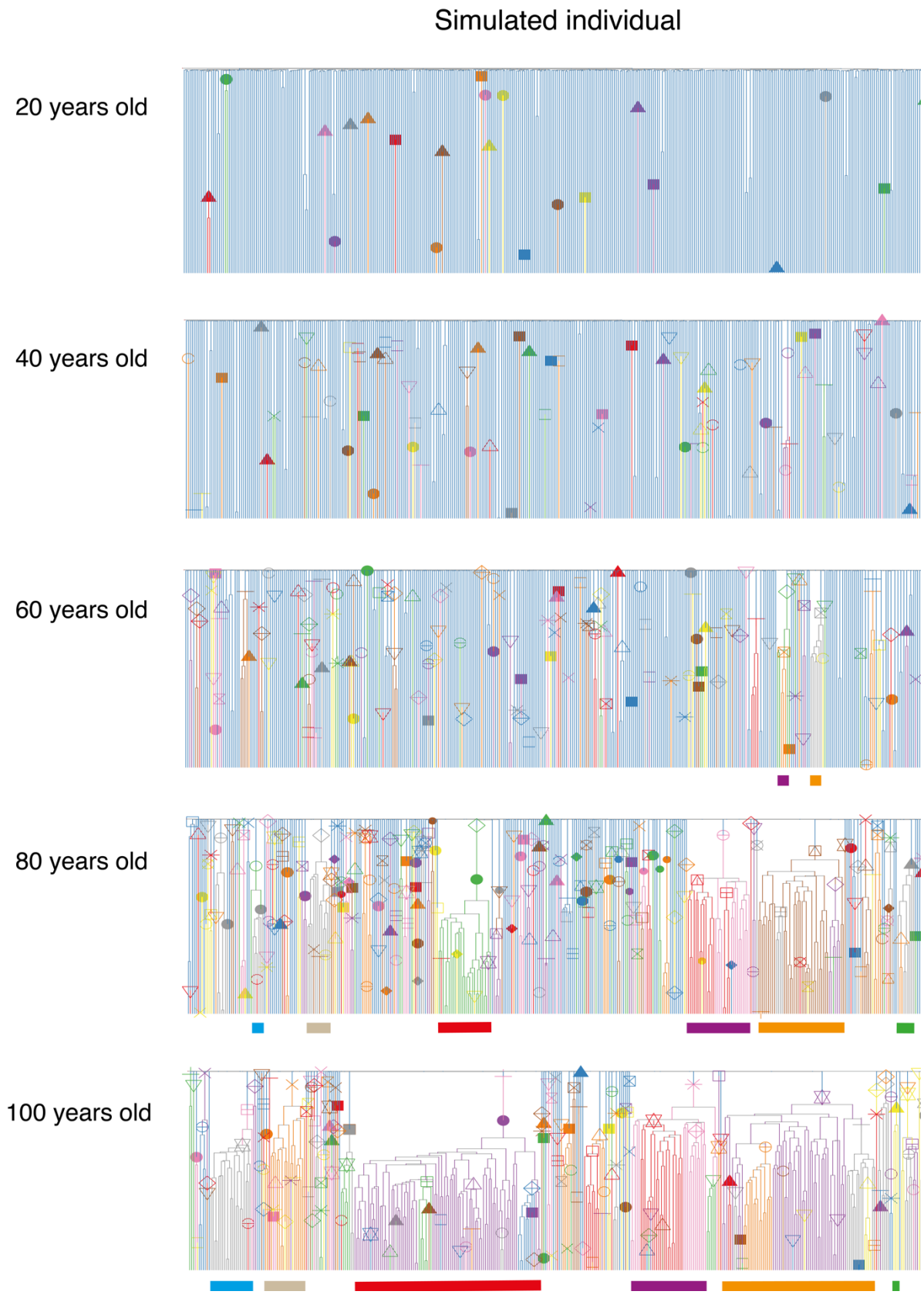


Appendix Fig. 4|Effect of population 'bottleneck'. a, Trajectories of $N\tau$ used as input to *rsimpop* for the simulations to create phylogenies in b. Note the Y axis depicting $N\tau$ is on a log scale. **b,** Phylogenies created by randomly sampling 380 cells from the final full simulated population of 100,000. Each simulation has an initial $N\tau$ of 100,000 with a decline to 10,000 during a period of midlife. In all cases $N\tau$ is the same as the population size (N), as the generation time (τ) is 1 year. The blue boxes indicate the period of time in which the population size is decreased. The *phylodyn* trajectories to the right of each simulated phylogeny use the pattern of coalescent events to recover the input trajectories for $N\tau$. The blue lines mark the times of change in $N\tau$.

Positive selection

Positive selection can also be simulated in the phylogenies as illustrated in **Fig. 4.12 and Appendix Fig. 5**. These figures show phylogenies drawn from HSC populations where *N* is 100,00 and τ is 1 year, with the population as a whole acquiring 200 driver mutations per year, although not all of these will be fixed in the population. The fitness effect of the driver mutations is drawn from a fitness effect gamma distribution (with shape = 0.47 and rate = 34) that incorporates a fitness effect threshold of 5% (**Fig. 4.13**). These parameters allow accurate recapitulation of the observed phylogenies across the human lifespan. The simulations illustrate how, although driver mutations are present in the phylogenies of individuals aged below 40, they do not typically impact the pattern of observed coalescences until later in life. This observation provides support for the accuracy of our estimates of $N\tau$ in the two youngest individuals in our cohort. In addition, the simulations demonstrate how large clones typically only become detectable after the age of 60, despite the founding driver mutations having been acquired decades earlier (typically in the first 3-4 decades of life). They also illustrate the range of older phylogenies (similar to the range of topologies in our real phylogenies) that can be generated from the stochastic process of driver acquisition. The simulations show how by age 115 years the haematopoietic system could commonly be sustained by just two clones with no known driver mutations, as has been previously reported in a single real individual[85].

The simple model we use predicts that by age 80, typically > 90% HSCs contain at least 1 driver mutation. In addition, there is a high prevalence of cells containing multiple drivers, such that in later life clonal competition between driver containing clones with different fitness effects can cause complex clonal dynamics. This is illustrated by that fact that some of the highlighted clades remain stable in size over the last few decades of life, while others may even decline in size. In all illustrated cases one or more 'fittest' clones continues to expand into extreme old age.



Appendix Fig. 5| Positive selection simulation for single individual. Phylogenies of 380 cells sampled from a population of 100,000 cells that has been maintained at a constant $N\tau$ over life, with incorporation of positively selected 'driver mutations'. The driver mutations have a fitness

effect > 5% (drawn from a gamma distribution with shape = 0.47 and rate = 34) and enter the population at a rate of 200 per year. These are the optimal estimates of these parameters based on our ABC modelling. The inclusion of these driver mutations is able to recapitulate a similar clade size distribution to that observed in the real HSPC phylogenies of the observed individuals across the whole age range. However, including driver mutations does not fully recapitulate the observed lack of coalescent events in the last 10-15 years of life, showing that an increase in $N\tau$ over this time is also required to fully recreate the patterns of coalescences in the real phylogenies. Driver mutations are marked with a symbol and their descendent clades are coloured. In all cases $N\tau$ is the same as the population size (N) as the generation time (τ) in all simulations is fixed at 1 year. The symbols / colours are not consistent for driver mutations between plots. The largest clades are therefore coloured in a consistent way beneath the plots to show how their size changes over time. The simulated phylogenies illustrate the complex clonal dynamics that can occur in later life as a result of clonal competition. While the majority of clades continue to expand, others stay relatively stable and some reduce in size. The phylogenies also show that by the age of 80 typically > 90% of HSCs in the population carry at least one driver mutation.

References

1. Nangalia, J., Mitchell, E. & Green, A. R. Clonal approaches to understanding the impact of mutations on hematologic disease development. *Blood* **133**, 1436–1445 (2019).
2. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *bioRxiv* 2021.08.16.456475 (2021) doi:10.1101/2021.08.16.456475.
3. Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* (2021) doi:10.1038/s41586-021-03548-6.
4. Williams, N. *et al.* Life histories of myeloproliferative neoplasms inferred from phylogenies. doi:10.1038/s41586-021-04312-6.
5. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).
6. Naik, S. H. *et al.* Diverse and heritable lineage imprinting of early haematopoietic progenitors. (2013) doi:10.1038/nature12013.
7. Lu, R., Neff, N. F., Quake, S. R., Weissman, I. L. & Author, N. B. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding HHS Public Access Author manuscript. (2012) doi:10.1038/nbt.1977.
8. Biasco, L. *et al.* In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. *Cell Stem Cell* **19**, 107–119 (2016).
9. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. (2018) doi:10.1038/nature25022.
10. Baum, C. M., Weissman, I. L., Tsukamoto, A. S., Buckle, A. M. & Peault, B. Isolation of a candidate human hematopoietic stem-cell population. *Proc. Natl. Acad. Sci.* **89**, 2804–2808 (1992).
11. Bhatia, M., Bonnet, D., Murdoch, B., Gan, O. I. & Dick, J. E. A newly discovered class of human hematopoietic cells with SCID- repopulating activity. *Nat. Med.* **4**, 1038–1045 (1998).
12. Lansdorp, P. M., Sutherland, H. J. & Eaves, C. J. Selective expression of CD45 isoforms on functional subpopulations of CD34+ hemopoietic cells from human bone marrow.

J Exp Med **172**, 363–366 (1990).

13. Notta, F. *et al.* Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. *Science (80-.).* **333**, 218–221 (2011).
14. Hwang, S. M. *et al.* Are clonal cells circulating in the peripheral blood of myelodysplastic syndrome?: Quantitative comparison between bone marrow and peripheral blood by targeted gene sequencing and fluorescence in situ hybridization. *Leukemia Research* vol. 71 92–94 (2018).
15. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
16. Nowell, P. & Hungerford, D. A minute chromosome in human chronic granulocytic leukemia. *Science (80-.).* 1497 (1960).
17. Rowley, J. D. Chromosomal translocations: revisited yet again. *Blood* **112**, 2183–2189 (2008).
18. Linder, D. & Gartler, S. M. Glucose-6-phosphate dehydrogenase mosaicism: utilization as a cell marker in the study of leiomyomas. *Science* **150**, 67–69 (1965).
19. Beutler, E., Collins, Z. & Irwin, L. E. Value of genetic variants of glucose-6-phosphate dehydrogenase in tracing the origin of malignant tumors. *N. Engl. J. Med.* **276**, 389–391 (1967).
20. Fialkow, P. J., Gartler, S. M. & Yoshida, A. Clonal origin of chronic myelocytic leukemia in man. *Proc Natl Acad Sci U S A* **58**, 1468–71 (1967).
21. Adamson, J. W., Fialkow, P. J., Murphy, S., Prchal, J. F. & Steinmann, L. Polycythemia vera: stem-cell and probable clonal origin of the disease. *N. Engl. J. Med.* **295**, 913–916 (1976).
22. Busque, L. *et al.* Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat. Genet.* **44**, 1179–1181 (2012).
23. Moore, M. a. S., Williams, N. & Metcalf, D. In Vitro Colony Formation by Normal and Leukemic Human Hematopoietic Cells: Characterization of the Colony-Forming Cells. *JNCI J. Natl. Cancer Inst.* **50**, 603–623 (1973).
24. Belluschi, S. *et al.* Myelo-lymphoid lineage restriction occurs in the human haematopoietic stem cell compartment before lymphoid-primed multipotent progenitors. doi:10.1038/s41467-018-06442-4.
25. Becker, A. J., McCULLOCH, E. A. & Till, J. E. Cytological demonstration of the clonal

nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* **197**, 452–454 (1963).

26. Osawa, M., Hanada, K., Hamada, H. & Nakauchi, H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* **273**, 242–245 (1996).
27. Goyama, S., Wunderlich, M. & Mulloy, J. C. Xenograft models for normal and malignant stem cells. *Blood* **125**, 2630–2640 (2015).
28. Glauche, I., Bystrykh, L., Eaves, C. & Roeder, I. Stem cell clonality - Theoretical concepts, experimental techniques, and clinical challenges. *Blood Cells, Mol. Dis.* **50**, 232–240 (2013).
29. Naldini, L., Blömer, U., Gage, F. H., Trono, D. & Verma, I. M. Efficient transfer, integration, and sustained long-term expression of the transgene in adult rat brains injected with a lentiviral vector. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 11382–11388 (1996).
30. Belderbos, M. E. *et al.* Clonal selection and asymmetric distribution of human leukemia in murine xenografts revealed by cellular barcoding. *Blood* **129**, 3210–3220 (2017).
31. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
32. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).
33. Xu, X. *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886–895 (2012).
34. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
35. Jones, O. R. *et al.* Diversity of ageing across the tree of life. *Nature* **505**, 169–173 (2014).
36. Edwards, R. D. & Tuljapurkar, S. Inequality in life spans and a new perspective on mortality convergence across industrialized countries. *Popul. Dev. Rev.* **31**, 645–674 (2005).
37. Burnet, F. M. Intrinsic mutagenesis: A genetic basis of ageing. *Pathology* **6**, 1–11 (1974).

38. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
39. Vijg, J. & Dong, X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell* vol. 182 12–23 (2020).
40. Armanios, M. *et al.* Short Telomeres are Sufficient to Cause the Degenerative Defects Associated with Aging. *Am. J. Hum. Genet.* **85**, 823–832 (2009).
41. Armanios, M. & Blackburn, E. H. The telomere syndromes. *Nature reviews. Genetics* vol. 13 693–704 (2012).
42. Rufer, N. *et al. Telomere Fluorescence Measurements in Granulocytes and T Lymphocyte Subsets Point to a High Turnover of Hematopoietic Stem Cells and Memory T Cells in Early Childhood*. *J. Exp. Med* vol. 190 http://www.jem.org (1999).
43. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, 115 (2013).
44. Flach, J. *et al.* Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. *Nature* **512**, 198–202 (2014).
45. Bogeska, R. *et al.* Hematopoietic stem cells fail to regenerate following inflammatory 1 challenge. *bioRxiv* 2020.08.01.230433 (2020) doi:10.1101/2020.08.01.230433.
46. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* vol. 153 1194–1217 (2013).
47. Kovtonyuk, L. V., Fritsch, K., Feng, X., Manz, M. G. & Takizawa, H. Inflamm-aging of hematopoiesis, hematopoietic stem cells, and the bone marrow microenvironment. *Frontiers in Immunology* vol. 7 (2016).
48. Mayack, S. R., Shadrach, J. L., Kim, F. S. & Wagers, A. J. Systemic signals regulate ageing and rejuvenation of blood stem cell niches. *Nat. 2010 4637280* **463**, 495–500 (2010).
49. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
50. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
51. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
52. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia.

Cell **150**, 264–278 (2012).

53. Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep.* **25**, 2308–2316 (2018).
54. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
55. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
56. Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science (80-.).* **370**, (2020).
57. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
58. Vaziri, H. *et al. Evidence for a mitotic clock in human hematopoietic stem cells: Loss of telomeric DNA with age*. *Proc. Nati. Acad. Sci. USA* vol. 91 (1994).
59. Hayflick, L. & Moorhead, P. S. The serial cultivation of human diploid cell strains. *Exp. Cell Res.* **25**, 585–621 (1961).
60. Palm, W. & De Lange, T. How Shelterin Protects Mammalian Telomeres. *http://dx.doi.org/10.1146/annurev.genet.41.110306.130350* **42**, 301–334 (2008).
61. Maciejowski, J., Li, Y., Bosco, N. & Campbell, P. J. Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell* **163**, 1641–1654 (2015).
62. Martinez, P. & Blasco, M. A. Role of shelterin in cancer and aging. *Aging Cell* **9**, 653–666 (2010).
63. Wimazal, F. *et al.* Idiopathic cytopenia of undetermined significance (ICUS) versus low risk MDS: The diagnostic interface. *Leuk. Res.* **31**, 1461–1468 (2007).
64. Valent, P. *et al.* Idiopathic cytopenia of undetermined significance (ICUS) and idiopathic dysplasia of uncertain significance (IDUS), and their distinction from low risk MDS. *Leuk. Res.* **36**, 1–5 (2012).
65. Sudo, K., Ema, H., Morita, Y. & Nakauchi, H. Age-Associated Characteristics of Murine Hematopoietic Stem Cells. *J. Exp. Med.* **192**, 1273–1280 (2000).
66. Rossi, D. J. *et al.* Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proc. Natl. Acad. Sci.* **102**, 9194–9199 (2005).
67. Dykstra, B., Olthof, S., Schreuder, J., Ritsema, M. & de Haan, G. Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells. *J. Exp. Med.*

208, 2691–2703 (2011).

68. Dykstra, B., Olthof, S., Schreuder, J., Ritsema, M. & Haan, G. De. Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells. *J. Exp. Med.* **208**, 2691–2703 (2011).
69. Beerman, I. Accumulation of DNA damage in the aged hematopoietic stem cell compartment. *Seminars in Hematology* vol. 54 12–18 (2017).
70. Mcmichael, A., Simon, A. K. & Hollander, G. A. Evolution of the immune system in humans from infancy to old age. doi:10.1098/rspb.2014.3085.
71. Beerman, I. *et al.* Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *Proc. Natl. Acad. Sci.* **107**, 5465–5470 (2010).
72. Pang, W. W. *et al.* Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20012–20017 (2011).
73. Yamamoto, R. *et al.* Large-Scale Clonal Analysis Resolves Aging of the Mouse Hematopoietic Stem Cell Compartment. *Cell Stem Cell* **22**, 600–607.e4 (2018).
74. Bernitz, J. M., Kim, H. S., MacArthur, B., Sieburg, H. & Moore, K. Hematopoietic Stem Cells Count and Remember Self-Renewal Divisions. *Cell* **167**, 1296–1309.e10 (2016).
75. Walter, D. *et al.* Exit from dormancy provokes DNA-damage-induced attrition in haematopoietic stem cells. *Nature* **520**, 549–552 (2015).
76. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
77. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
78. McKerrell, T. *et al.* Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
79. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* (2014) doi:10.1038/nm.3733.
80. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, (2016).
81. McKerrell, T. *et al.* Leukemia-Associated Somatic Mutations Drive Distinct Patterns of

Age-Related Clonal Hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).

82. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
83. Pich, O., Reyes-Salazar, I., Gonzalez-Perez, A. & Lopez-Bigas, N. Discovering the drivers of clonal hematopoiesis. *bioRxiv* 2020.10.22.350140 (2020) doi:10.1101/2020.10.22.350140.
84. Poon, G. Y. P., Watson, C. J., Fisher, D. S. & Blundell, J. R. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. *Nat. Genet.* **53**, 1597–1605 (2021).
85. Holstege, H. *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res.* **24**, 733–742 (2014).
86. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
87. de Kanter, J. K. *et al.* Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. *Cell Stem Cell* **28**, 1726–1739.e6 (2021).
88. Catlin, S. N., Busque, L., Gale, R. E., Guttorp, P. & Abkowitz, J. L. The replication rate of human hematopoietic stem cells in vivo. *Blood* **117**, 4460–4466 (2011).
89. Werner, B. *et al.* Reconstructing the in vivo dynamics of hematopoietic stem cells from telomere length distributions. *Elife* **4**, (2015).
90. Lan, S., Palacios, J. A., Karcher, M., Minin, V. N. & Shahbaba, B. An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics* **31**, 3282–3289 (2015).
91. Mimasaka, S. Postmortem cytokine levels and the cause of death. *Tohoku J. Exp. Med.* **197**, 145–150 (2002).
92. Schwarz, P. *et al.* Brain Death-Induced Inflammatory Activity is Similar to Sepsis-Induced Cytokine Release. *Cell Transplant.* **27**, 1417–1424 (2018).
93. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
94. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1744 (2018).
95. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents.

Cell **177**, 821–836.e16 (2019).

96. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, (2014).
97. Rustad, E. H. *et al.* Timing the initiation of multiple myeloma. *Nat. Commun. 2020 111* **11**, 1–14 (2020).
98. Maura, F. *et al.* The mutagenic impact of melphalan in multiple myeloma. *Leuk. 2021 358* **35**, 2145–2150 (2021).
99. Pich, O. *et al.* The mutational footprints of cancer therapies. doi:10.1101/683268.
100. Robinson, P. S. *et al.* Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat. Genet.* doi:10.1038/s41588-021-00930-y.
101. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, (2019).
102. Coorens, T. H. H. *et al.* Clonal hematopoiesis and therapy-related myeloid neoplasms following neuroblastoma treatment. *Blood* **137**, 2992–2997 (2021).
103. Schoedel, K. B. *et al.* The bulk of the hematopoietic stem cell population is dispensable for murine steady-state and stress hematopoiesis. *Blood* **128**, 2285–2296 (2016).
104. Randall, T. D. & Weissman, I. L. Phenotypic and functional changes induced at the clonal level in hematopoietic stem cells after 5-fluorouracil treatment. *Blood* **89**, 3596–3606 (1997).
105. Busch, K. *et al.* Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* **518**, 542–546 (2015).
106. Bolton, K. L. *et al.* Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nat. Genet.* **52**, 12219–1226 (2020).
107. Hsu, J. I. *et al.* PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. *Cell Stem Cell* **23**, 700 (2018).
108. Coombs, C. C. *et al.* Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and impacts clinical outcome. *Cell Stem Cell* **21**, 374 (2017).
109. Kollmannsberger, C., Hartmann, J. T., Kanz, L. & Bokemeyer, C. Risk of secondary myeloid leukemia and myelodysplastic syndrome following standard-dose chemotherapy or high-dose chemotherapy with stem cell support in patients with

potentially curable malignancies. *J. Cancer Res. Clin. Oncol.* **124**, 207–214 (1998).

110. Notta, F. *et al.* Isolation of Single Human Hematopoietic Stem Cells Capable of Long-Term Multilineage Engraftment. *Science (80-.).* **333**, 218–221 (2011).
111. Huntsman, H. D. *et al.* Human hematopoietic stem cells from mobilized peripheral blood can be purified based on CD49f integrin expression. *Blood* **126**, 1631–1633 (2015).
112. Laurenti, E. *et al.* CDK6 Levels Regulate Quiescence Exit in Human Hematopoietic Stem Cells. *Cell Stem Cell* **16**, 302 (2015).
113. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science (80-.).* **351**, (2016).
114. Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* doi:10.1038/s41596-020-00437-6.
115. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, (2020).
116. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. doi:10.1002/cpbi.20.
117. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. doi:10.1002/0471250953.bi1507s52.
118. Tim H Coorens, A. H. *et al.* TITLE Extensive phylogenies of human development reveal variable embryonic patterns. doi:10.1101/2020.11.25.397828.
119. Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* 1–6 (2021) doi:10.1038/s41586-021-03548-6.
120. Cameron, D. L. *et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. (2017) doi:10.1101/gr.222109.117.
121. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16910–16915 (2010).
122. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149**, 994–1007 (2012).
123. Lee-Six, H. & Kent, D. G. Tracking hematopoietic stem cells and their progeny using whole-genome sequencing. *Exp. Hematol.* **83**, 12–24 (2020).
124. Farmery, J. H. R. *et al.* Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci. Rep.* **8**, (2018).

125. Frenck, R. W., Blackburn, E. H. & Shannon, K. M. *The rate of telomere sequence loss in human leukocytes varies with age*. *Cell Biology* vol. 95 www.pnas.org. (1998).
126. Thi Hoang, D. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. doi:10.1186/s12862-018-1131-3.
127. Williams, N. *et al.* Phylogenetic reconstruction of myeloproliferative neoplasm reveals very early origins and lifelong evolution. doi:10.1101/2020.11.09.374710.
128. Williams, N. *et al.* Life histories of myeloproliferative neoplasms inferred from phylogenies. | *Nat.* | **602**, (2022).
129. Williams, N. *et al.* Phylogenetic reconstruction of myeloproliferative neoplasm reveals very early origins and lifelong evolution. doi:10.1101/2020.11.09.374710.
130. Csilléry, K., François, O. & Blum, M. G. B. Abc: An R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479 (2012).
131. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
132. Gelman, A. *et al. Bayesian data analysis*. (Chapman and Hall, CRC Press, 2004).
133. Fabre, M. A. *et al.* The longitudinal dynamics and natural history of clonal haematopoiesis. *bioRxiv* 2021.08.12.455048 (2021) doi:10.1101/2021.08.12.455048.
134. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
135. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
136. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
137. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
138. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
139. Vaser, R., Adusumalli, S., Ngak Leng, S., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* (2015) doi:10.1038/nprot.2015.123.
140. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. doi:10.1002/0471142905.hg0720s76.
141. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous

and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).

142. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173**, 2187–2198 (2006).
143. Watson, C. J. *et al.* The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science (80-.).* **367**, 1449–1454 (2020).
144. Lan, S., Palacios, J. A., Karcher, M., Minin, V. N. & Shahbaba, B. An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics* **31**, 3282–3289 (2015).
145. Duncavage, E. J. *et al.* Genome Sequencing as an Alternative to Cytogenetic Analysis in Myeloid Cancers. *N. Engl. J. Med.* **384**, 924–935 (2021).
146. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
147. Klco, J. M. *et al.* Association between mutation clearance after induction therapy and outcomes in acute myeloid leukemia. *JAMA - J. Am. Med. Assoc.* **314**, 811–822 (2015).
148. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
149. Mackey, M. C. Cell kinetic status of haematopoietic stem cells. *Cell Prolif.* **34**, 71 (2001).
150. Shepherd, B. E., Guttorp, P., Lansdorp, P. M. & Abkowitz, J. L. *Estimating human hematopoietic stem cell kinetics using granulocyte telomere lengths*. *Experimental Hematology* vol. 32 (2004).
151. Abkowitz, J. L., Catlin, S. N., McCallie, M. T. & Guttorp, P. Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. *Blood* **100**, 2665–7 (2002).
152. Watson, C. J. *et al.* The evolutionary dynamics and fitness landscape of clonal haematopoiesis. *bioRxiv* 569566 (2019) doi:10.1101/569566.
153. Bradford, G. B., Williams, B., Rossi, R. & Bertoncello, I. Quiescence, cycling, and turnover in the primitive hematopoietic stem cell compartment. *Exp. Hematol.* **25**, 445–453 (1997).
154. Cheshier, S. H., Morrison, S. J., Liao, X. & Weissman, I. L. In vivo proliferation and cell cycle kinetics of long-term self-renewing hematopoietic stem cells. *Proc. Natl. Acad.*

Sci. **96**, 3120–3125 (1999).

155. Kaschutnig, P. *et al.* The Fanconi anemia pathway is required for efficient repair of stress-induced DNA damage in haematopoietic stem cells. *Cell Cycle* **14**, 2734–2742 (2015).
156. Foudi, A. *et al.* Analysis of histone 2B-GFP retention reveals slowly cycling hematopoietic stem cells. *Nat. Biotechnol.* **27**, 84–90 (2009).
157. Dykstra, B. *et al.* Long-Term Propagation of Distinct Hematopoietic Differentiation Programs In Vivo. *Cell Stem Cell* **1**, 218–229 (2007).
158. Ho, T. T. *et al.* Aged hematopoietic stem cells are refractory to bloodborne systemic rejuvenation interventions. (2021) doi:10.1084/jem.20210223.
159. Schoenmakers, E. *et al.* Mutations in the selenocysteine insertion sequence-binding protein 2 gene lead to a multisystem selenoprotein deficiency disorder in humans. *J. Clin. Invest.* **120**, 4220–4235 (2010).
160. Danielsson, M. *et al.* Longitudinal changes in the frequency of mosaic chromosome Y loss in peripheral blood cells of aging men varies profoundly between individuals. *Eur. J. Hum. Genet.* **28**, 349–357 (2020).
161. Aubert, G., Baerlocher, G. M., Vulto, I. ¤, Poon, S. S. & Lansdorp, P. M. Collapse of Telomere Homeostasis in Hematopoietic Cells Caused by Heterozygous Mutations in Telomerase Genes. *PLoS Genet* **8**, 1002696 (2012).
162. Baerlocher, G. M., Rice, K., Vulto, I. & Lansdorp, P. M. Longitudinal data on telomere length in leukocytes from newborn baboons support a marked drop in stem cell turnover around 1 year of age. *Aging Cell* **6**, 121–123 (2007).
163. Wilson, A. *et al.* Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. *Cell* **135**, 1118–1129 (2008).
164. HARTSOCK, R. J., SMITH, E. B. & PETTY, C. S. Normal Variations with Aging of the Amount of Hematopoietic Tissue in Bone Marrow from the Anterior Iliac Crest: A Study Made from 177 Cases of Sudden Death Examined by Necropsy. *Am. J. Clin. Pathol.* **43**, 326–331 (1965).
165. Ricci, C. *et al.* Normal age-related patterns of cellular and fatty bone marrow distribution in the axial skeleton: MR imaging study. *https://doi.org/10.1148/radiology.177.1.2399343* **177**, 83–88 (1990).
166. Kuranda, K. *et al.* Age-related changes in human hematopoietic stem/progenitor cells.

Aging Cell **10**, 542–546 (2011).

167. Mende, N. *et al.* Quantitative and molecular differences distinguish adult human medullary and extramedullary haematopoietic stem and progenitor cell landscapes. *bioRxiv* 2020.01.26.919753 (2020) doi:10.1101/2020.01.26.919753.
168. Karcher, M. D., Palacios, J. A., Lan, S. & Minin, V. N. *phylodyn: an R package for phylodynamic simulation and inference*. https://github.com/mdkarcher/phylodyn.
169. Karcher, M. D., Palacios, J. A., Lan, S. & Minin, V. N. phylodyn: an R package for phylodynamic simulation and inference. doi:10.1111/1755-0998.12630.
170. Barile, M. *et al.* Hematopoietic stem cells self-renew symmetrically or gradually proceed to differentiation. doi:10.1101/2020.08.06.239186.
171. Ito, K. *et al.* Self-renewal of a purified Tie2+ hematopoietic stem cell population relies on mitochondrial clearance. *Science (80-.).* **354**, 1156–1160 (2016).
172. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
173. De Haan, G. & Lazare, S. S. Aging of hematopoietic stem cells. *Blood* vol. 131 479–487 (2018).
174. Pang, W. W., Schrier, S. L. & Weissman, I. L. Age-associated changes in human hematopoietic stem cells. *Seminars in Hematology* vol. 54 39–42 (2017).
175. Martincorena, I., Raine, K. M., Davies, H., Stratton, M. R. & Campbell, P. J. Universal Patterns of Selection in Cancer and Somatic Tissues. (2017) doi:10.1016/j.cell.2017.09.042.
176. Van Egeren, D. *et al.* Reconstructing the Lineage Histories and Differentiation Trajectories of Individual Cancer Cells in Myeloproliferative Neoplasms. *Cell Stem Cell* **28**, 514–523.e9 (2021).
177. Wong, T. N. *et al.* Cellular stressors contribute to the expansion of hematopoietic clones of varying leukemic potential. *Nat. Commun.* **9**, (2018).
178. Beauchamp, E. M. *et al.* ZBTB33 Is Mutated in Clonal Hematopoiesis and Myelodysplastic Syndromes and Impacts RNA Splicing. *Blood Cancer Discov.* **2**, 500 (2021).
179. Zhu, Y., Wang, G. Z., Cingöz, O. & Goff, S. P. NP220 mediates silencing of unintegrated retroviral DNA. *Nature* (2018) doi:10.1038/s41586-018-0750-6.
180. Douse, C. H. *et al.* TASOR is a pseudo-PARP that directs HUSH complex assembly and epigenetic transposon control. *Nat. Commun. 2020 111* **11**, 1–16 (2020).

181. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* vol. 128 693–705 (2007).
182. Gozdecka, M. *et al.* UTX-mediated enhancer and chromatin remodeling suppresses myeloid leukemogenesis through noncatalytic inverse regulation of ETS and GATA programs. *Nat. Genet.* **50**, 883–894 (2018).
183. Poon, G., Watson, C. J., Fisher, D. S. & Blundell, J. R. Synonymous mutations reveal genome-wide driver mutation rates in healthy tissues. *bioRxiv* 2020.10.08.331405 (2020) doi:10.1101/2020.10.08.331405.
184. Izzo, F. *et al.* DNA methylation disruption reshapes the hematopoietic differentiation landscape. *Nat. Genet.* **52**, 378–387 (2020).
185. Nam, A. S. *et al.* Single-cell multi-omics of human clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation. *bioRxiv* 2022.01.14.476225 (2022) doi:10.1101/2022.01.14.476225.
186. Bhatia, S. Therapy-related myelodysplasia and acute myeloid leukemia. *Semin Oncol* **40**, (2013).
187. Morton, L. M. *et al.* Association of Chemotherapy for Solid Tumors With Development of Therapy-Related Myelodysplastic Syndrome or Acute Myeloid Leukemia in the Modern Era. *JAMA Oncol.* **5**, 318 (2019).
188. Groopman, J. E. & Itri, L. M. Chemotherapy-induced anemia in adults: Incidence and treatment. *Journal of the National Cancer Institute* vol. 91 1616–1634 (1999).
189. Crawford, J., Dale, D. C. & Lyman, G. H. Chemotherapy-Induced Neutropenia: Risks, Consequences, and New Directions for Its Management. *Cancer* vol. 100 228–237 (2004).
190. Shaw Mph, J. L. *et al.* | INTRODUC TI ON The incidence of thrombocytopenia in adult patients receiving chemotherapy for solid tumors or hematologic malignancies. (2021) doi:10.1111/ejh.13595.
191. Szász, R., Telek, B. & Illés, Á. Fludarabine-Cyclophosphamide-Rituximab Treatment in Chronic Lymphocytic Leukemia, Focusing on Long Term Cytopenias Before and After the Era of Targeted Therapies. *Pathol. Oncol. Res.* **27**, 97 (2021).
192. Zan, H. *et al.* The Translesion DNA Polymerase ζ Plays a Major Role in Ig and bcl-6 Somatic Hypermutation. *Immunity* **14**, 643 (2001).
193. Martin, S. K. & Wood, R. D. DNA polymerase ζ in DNA replication and repair. *Nucleic*

Acids Res. **47**, 8348–8361 (2019).

194. Plosky, B. S. & Woodgate, R. Switching from high-fidelity replicases to low-fidelity lesion-bypass polymerases. *Curr. Opin. Genet. Dev.* **14**, 113–119 (2004).
195. Wolf, J. *et al.* Peripheral Blood Mononuclear Cells of a Patient With Advanced Hodgkin's Lymphoma Give Rise to Permanently Growing Hodgkin-Reed Sternberg Cells. *Blood* **87**, 3418–3428 (1996).
196. Kaplanis, J. *et al.* Genetic and pharmacological causes of germline hypermutation. *bioRxiv* 2021.06.01.446180 (2021) doi:10.1101/2021.06.01.446180.
197. Zicchedduu, B. *et al.* Integrative analysis of the genomic and transcriptomic landscape of double-refractory multiple myeloma. *Blood Adv.* **4**, 830–844 (2020).
198. Landau, H. J. *et al.* Accelerated single cell seeding in relapsed multiple myeloma. doi:10.1038/s41467-020-17459-z.
199. Boot, A. *et al.* In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res.* **28**, 654–665 (2018).
200. Mehmood, R. K. Review of Cisplatin and Oxaliplatin in Current Immunogenic and Monoclonal Antibody Treatments. *Oncol. Rev.* **8**, (2014).
201. O'dwyer, P. J., Stevenson, J. P. & Johnson, S. W. Clinical Pharmacokinetics and Administration of Established Platinum Drugs. *Drugs* **59**, 19–27 (2000).
202. Zhang, W., Gou, P., Dupret, J. M., Chomienne, C. & Rodrigues-Lima, F. Etoposide, an anticancer drug involved in therapy-related secondary leukemia: Enzymes at play. *Transl. Oncol.* **14**, 101169 (2021).
203. Nagel, C. I. *et al.* Effect of chemotherapy delays and dose reductions on progression free and overall survival in the treatment of epithelial ovarian cancer. *Gynecol. Oncol.* **124**, 221–224 (2012).
204. Liutkauskiene, S. *et al.* Retrospective analysis of the impact of anthracycline dose reduction and chemotherapy delays on the outcomes of early breast cancer molecular subtypes. *BMC Cancer* **18**, (2018).
205. Dawoud, A. A. Z., Tapper, W. J., Nicholas, • & Cross, C. P. Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking. *Leukemia* **34**, 2660–2672 (2020).
206. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood

DNA Sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).

207. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
208. Bertorelle, G., Benazzo, A. & Mona, S. ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Mol. Ecol.* **19**, 2609–2625 (2010).
209. Ho, T. T. *et al.* Aged hematopoietic stem cells are refractory to bloodborne systemic rejuvenation interventions. *J. Exp. Med.* **218**, (2021).
210. Meisel, M. *et al.* Microbial signals drive pre-leukaemic myeloproliferation in a Tet2-deficient host. *Nat. 2018 5577706* **557**, 580–584 (2018).
211. Hormaechea-Agulla, D. *et al.* Chronic infection drives Dnmt3a-loss-of-function clonal hematopoiesis via IFNγ signaling. *Cell Stem Cell* **28**, 1428–1442.e6 (2021).
212. Ramdas, B. *et al.* Driver Mutations in Leukemia Promote Disease Pathogenesis through a Combination of Cell-Autonomous and Niche Modulation. *Stem Cell Reports* **15**, 95–109 (2020).
213. Alex Cagan, A. *et al.* Somatic mutation rates scale with lifespan across mammals. doi:10.1101/2021.08.19.456982.
214. Abkowitz, J. L., Catlin, S. N., Mccallie, M. T. & Guttorp, P. Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. **100**, 2665–2667 (2002).
215. Cosgrove, J., Hustin, L. S. P., de Boer, R. J. & Perié, L. Hematopoiesis in numbers. *Trends Immunol.* **42**, 1100–1112 (2021).
216. Wilson, A. *et al.* Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. *Cell* **135**, 1118–1129 (2008).
217. Ganuza, M. *et al.* The global clonal complexity of the murine blood system declines throughout life and after serial transplantation Short title (48 characters including spaces): The clonal complexity of blood declines with age. *Blood First Ed. Pap.* (2019) doi:10.1182/blood-2018-09-873059.
218. Yu, K. R. *et al.* The impact of aging on primate hematopoiesis as interrogated by clonal tracking. *Blood* **131**, 1195–1205 (2018).
219. Li, R. *et al.* A body map of somatic mutagenesis in morphologically normal human tissues. *Nature* **597**, 398–403 (2021).

220. Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *538* | *Nat.* | **574**, (2019).
221. Olafsson, S. *et al.* Somatic Evolution in Non-neoplastic IBD-Affected Colon. *Cell* **182**, 672–684.e11 (2020).
222. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (80-.).* **348**, 880–886 (2015).
223. Fowler, J. C. *et al.* Selection of oncogenic mutant clones in normal human skin varies with body site. *Cancer Discov.* **11**, 340–361 (2021).
224. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
225. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science (80-.).* **362**, 911–917 (2018).
226. Duncavage, E. J. *et al.* Genome Sequencing as an Alternative to Cytogenetic Analysis in Myeloid Cancers. *N. Engl. J. Med.* **384**, 924–935 (2021).
227. Vassalli, G. Aldehyde dehydrogenases: Not just markers, but functional regulators of stem cells. *Stem Cells International* vol. 2019 (2019).
228. Kushner, B. H. *et al.* Reduced risk of secondary leukemia with fewer cycles of dose-intensive induction chemotherapy in patients with neuroblastoma. *Pediatr. Blood Cancer* **53**, 17–22 (2009).