The evolutionary dynamics of clonal haematopoiesis and its progression to acute myeloid leukaemia

Caroline Watson

Department of Oncology
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Corpus Christi College

December 2021

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 relevant Degree Committee.

Caroline Watson
December 2021

The evolutionary dynamics of clonal haematopoiesis and its progression to acute myeloid leukaemia

Caroline Watson

Abstract

Acute myeloid leukaemia (AML) is an aggressive blood cancer which claims the lives of 70-80% of patients within 5 years of diagnosis. Like many other cancers, AML usually develops as a consequence of serial acquisition of somatic driver mutations; a process that starts many years, or even decades, before diagnosis. This raises the prospect that early detection of 'pre-leukaemic' mutations could be used to identify individuals at high risk of developing AML, in whom early intervention could halt the disease before it fully develops. One of the difficulties with early detection of AML is that clonally expanded leukaemia-associated mutations are also found in the blood of healthy individuals, a phenomenon termed 'clonal haematopoiesis'. However, most individuals with clonal haematopoiesis will never progress to AML and so a key challenge is the identification of individuals most at risk. To do this, we need a better understanding of the evolutionary dynamics of clonal haematopoiesis in the years, or decades, before AML occurs and how this differs from the dynamics of clonal haematopoiesis in individuals that remain cancer-free.

We sought to understand this process by first studying the acquisition and expansion of the initial clonal haematopoiesis driver mutation. Using blood sequencing data amassed from ∼50,000 individuals, combined with insights from evolutionary theory, we developed a framework to quantify the mutation rates and fitness effects of clonal haematopoiesis variants down to single nucleotide resolution. This enabled us to build a league table of the fittest and potentially most pathogenic variants in blood. We also quantified the distribution of fitness across key clonal haematopoiesis genes and found the distribution to be highly skewed, with most mutations in these genes conferring either a weak or no fitness effect. Our framework also reveals that whilst cell-extrinsic effects are likely crucial in some situations, the combined effects of chance (when a mutation arises) and cell-intrinsic fitness differences are the major forces shaping clonal haematopoiesis.

Mosaic chromosomal alterations (mCAs) can also be important drivers in AML and ∼3% of individuals aged ∼40-70 have a clonally expanded mCA detectable in >1% of their blood cells. We therefore adapted our framework to quantify the mutation rates and fitness effects of mCAs in blood and applied this to data generated from ∼500,000 individuals in UK Biobank. We find most mCAs confer growth rates of ∼10-20% per year and find correlation between mCA fitness and blood cancer risk. In contrast to the strong age dependence observed in single nucleotide variant prevalence in blood, we find mCA age dependence to be more variable, particularly in women, suggesting the risk of acquisition and/ or expansion of certain mCAs is non-uniform throughout life and is influenced by gender-specific factors.

To determine how the dynamics of clonal haematopoiesis differs in individuals who progress to AML, we identified longitudinal blood samples that had been collected annually at multiple timepoints from individuals who subsequently developed AML, as well as age-matched controls who remained cancer free. We developed a custom error-corrected duplex sequencing platform to detect mutations in 34 clonal haematopoiesis/AML-associated genes, genome-wide mCAs and AML-associated translocations and used this to perform an integrative assessment of the genetic changes in these samples. We find there are four main evolutionary patterns in the years preceding AML diagnosis: linear evolution, evolution with clonal interference, static evolution and late evolution. We calculate the age at acquisition of the first and second mutations and, whilst the initial driver mutation is often acquired early in life, there are some very fit 'uber drivers' which appear to occur as the initial event just ∼4 years pre-diagnosis. We find that the variants we identified as 'highly fit' in clonal haematopoiesis are significantly enriched pre-AML and we were able to determine how fitness effects changed with the acquisition of subsequent mutations. NPM1 mutations, which characteristically occur late in AML development and have never been seen in individuals who do not progress to AML, can be detected as early as 2 years pre-AML diagnosis, highlighting the benefit afforded by low VAF variant calling, particularly in high-risk individuals.

This quantitative analysis of clonal haematopoiesis, combined with an integrated assessment of genetic changes in longitudinal blood samples from individuals who progress to AML, reveals important insights into the evolutionary dynamics of mutations in the years preceding AML. Understanding which features distinguish pre-malignant from benign clonal evolution is key for risk stratification of clonal haematopoiesis and will aid in the development of rational monitoring approaches and identification of those who may benefit from early intervention studies.

I would like to dedicate this thesis to all the AML patients and their families who I have had the privilege to know and care for.

Every drop on a Kaplan-Meier curve is somebody's loved one and each line whispers 'we must do better'.[†]

[†] Adapted from American Society of Haematology Annual Meeting 2018 (unknown author)

Acknowledgements

Foremost, I would like to express my sincere gratitude to my PhD supervisor Dr Jamie Blundell for his fantastic support, direction and mentorship over the past four years. His enthusiasm for science and his kindness, humour and unfaltering patience have made the past four years the most enjoyable of my life. It was an honour to be his first PhD student and I cannot imagine having had a better advisor. I would also like to thank all the past and present members of the Blundell lab for their friendship and support. I cannot think of a better group of people with whom I would have wanted to spend this journey.

I would like to thank and acknowledge our collaborators Prof Daniel Fisher and Alana Papula, from Stanford University, who worked with us on the development of the models used in the analysis of the clonal haematopoiesis datasets. Their knowledge and scientific rigour were inspiring. I would also like to thank Prof Todd Druley, Dr Andrew Young and Dr Wing Hing Wong, from Washington University in St Louis, for their openness in sharing pre-publication data with us. I would like to thank our collaborators at UCL, Prof Usha Menon and Dr Sophia Apostolidou, for sharing the UKCTOCS longitudinal blood samples with us. Dr Sophia Apostolidou often went above and beyond in helping us with this project, for which I am incredibly grateful. A special thanks to Wendy Alderton, without whom we would not have been aware these incredible samples even existed.

I would like to express my appreciation to the University of Cambridge and the CRUK Cambridge Centre Early Detection programme for giving me the opportunity to do this research. This work was funded by CRUK, UK Research and Innovation and The Henry Lumley Charitable Trust and I am very grateful to them all for their financial assistance.

On a personal note I would like to thank Andy for his kindness, support and tolerance of my late night keyboard tapping. Finally, but by no means least, I am eternally grateful to my parents and brother for their unfailing love, support and encouragement, without which it is unlikely I would be here today. My parents have always read everything I have written and this thesis has been no exception. I hope they are as proud of it as I am.

Contents

List of Figures

List of Tables

1

Introduction

1.1 Cancer as an evolutionary process

Cancer develops as a result of an evolutionary process that involves the stepwise accumulation of genetic or epigenetic changes within a cell over time[1,2] . The acquisition of somatic mutations with age is ubiquitous in healthy tissues such as the blood[3–9], skin[10–12] and gastrointestinal tract[13,14]. Whilst many of these mutations will be detrimental to the cell (deleterious mutations, e.g. a frame-shift mutation in a gene essential for DNA replication) or of no functional consequence (neutral mutations, e.g. synonymous mutations), some mutations confer a fitness advantage (beneficial mutations) by promoting cell proliferation and/or cell survival[15]. If these beneficial mutations occur in cells with an inherent, or acquired, capacity for self-renewal (e.g. stem cells), the mutation is propagated through clonal expansion[16], which increases the likelihood of further mutations being acquired and may ultimately result in the clone becoming cancerous[6–8]. This raises the possibility that cells harbouring early pre-cancerous mutations may be detectable in individuals in the years, or decades, before they show signs of the disease.

1.2 Acute myeloid leukaemia as an evolutionary process

Acute myeloid leukaemia (AML) is an aggressive blood cancer, characterised by clonal expansion and uncontrolled proliferation of abnormal undifferentiated myeloid precursor cells in the bone marrow, which results in impaired haematopoiesis and bone marrow failure[17,18]. AML incidence rates rise steadily until age ∼50 and then more steeply from age ∼60, with peak incidence between ages 85-89[19]. Whole genome sequencing has revealed AML to be a genetically heterogeneous disease[13–15] with $>$ 5000 possible driver mutations (across ∼76 genes or genomic regions)[20], although an individual AML is associated with an average of only ∼5 ± 3 (SD) mutations in recurrently mutated genes[21]. Single-cell analyses have demonstrated that these mutations are serially acquired in haematopoietic stem cells (HSCs)[17] which, with their long lifespan and capacity for self-renewal[22] makes them vulnerable to the accumulation of mutations over time[3–9]. It is also possible that the mutations are serially acquired in more downstream multi-potent progenitors, but only if the cell is capable of self-renewal (either inherently, or due to the effect of the first mutation), otherwise the mutation will be lost due to terminal differentiation[23]. Mutations affecting epigenetic regulators (e.g. DNMT3A, TET2, ASXL1) have been shown to occur as early events in the step-wise accumulation of mutations pre-AML, with mutations conferring a proliferative advantage (e.g. FLT3 and NPM1) occurring late[21,24–26] (Figure 1.1). It follows that we should expect to find evidence of early ‘pre-leukaemic’ mutations in the blood of individuals before they develop *bona fide* AML, as well as in the blood of ‘healthy ageing’ individuals who do not acquire the full complement of mutations required to develop the disease.

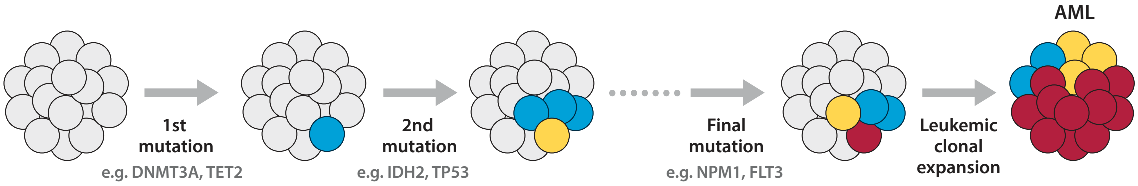


Figure 1.1 Step-wise accumulation of mutations in HSCs, leading to AML. Figure adapted from Jan et al[17].

1.3 Clonal haematopoiesis

Early evidence of clonal expansions in the blood ('clonal haematopoiesis') of healthy individuals came from the observation of a skewed pattern of X-chromosome inactivation in the peripheral blood cells of ∼20% of healthy women >65 years old[27]. The skewing was seen in all haematopoietic lineages, consistent with an HSC origin and, following the advent of next-generation sequencing (NGS), some cases were found to be associated with mutations in TET2[28], a gene recurrently mutated in AML[20,21].

In 2014, three large population-based studies[3–5] showed that the prevalence of clonal haematopoiesis, in individuals unselected for haematological malignancy, increases roughly exponentially with age, with somatic mutations detectable at >2% variant allele frequency (VAF) (i.e. >4% of blood cells) in 10% of individuals >65 years old[4]. The most commonly mutated genes were those found to be recurrently mutated in early pre-leukaemic stem cells[24], namely DNMT3A, TET2 and ASXL1, whereas NPM1 and FLT3-ITD mutations, which occur late in the evolution to AML[24–26], were not observed. This finding is consistent with the notion that mutations in DNMT3A, ASXL1 and TET2 act as 'initiating' mutations in pre-leukaemic cells, but that a 'co-operating' mutation in e.g. FLT3 or NPM1, is then required for leukaemic transformation[4,20]. Indeed, these large population-based studies showed that the presence of clonal haematopoiesis was a strong risk factor for the subsequent development of haematological malignancy (hazard ratio 11.1-12.9), with a risk of progression of ∼0.5-1% per year[3,29]. Bone marrow biopsy analysis of two individuals who subsequently developed AML confirmed that their cancers had arisen from their earlier detected clones[4]. More recent studies, using error-corrected sequencing methods, with the ability to detect variants at frequencies as low as 0.03%, have shown that clonal haematopoiesis is even more prevalent than originally thought, with 95% of individuals >60 years old having $\geq$1 somatic variant detectable in their peripheral blood[7]. Whilst the majority of studies have focused on gene mutations in clonal haematopoiesis, mosaic chromosomal alterations (mCAs) can also be found in the blood of healthy individuals[30–33], with recent studies showing that ∼3.5% of all individuals aged 40-70 years old have a clonally expanded mCA detectable in >1% of their blood cells[34,35]. Certain mCAs are associated with an increased risk of developing myeloid and/ or lymphoid malignancies, with an annual incidence of ∼0.5-1% per year[36]. This risk is significantly increased if clonal haematopoiesis variants are also present (hazard ratio ∼103).

1.4 Progression of clonal haematopoiesis to AML

AML affects ∼4 in 100,000 individuals per year in the UK and, despite recent advances in treatment, <30% survive for >5 years after their diagnosis[19]. The 'holy grail' would be to have the ability to reliably predict which individuals with clonal haematopoiesis are most at risk of developing AML and then intervene with targeted therapy to stop AML in its tracks before it develops.

How the mutational landscape differs in individuals who progress to AML was explored in recent studies, using blood samples collected from up to 3 timepoints pre-AML diagnosis[37–39]. These studies highlighted several key features associated with increased risk which were detectable in the blood ∼10 years prior to AML diagnosis. Pre-AML cases showed enrichment for mutations in TP53 (OR 47.2), IDH1/2 (OR 28.5), spliceosome genes (OR 7.4), TET2 (OR 5.8) and DNMT3A (OR 2.6)[37,38]. VAFs tended to be higher in pre-AML cases, with ∼40% having a mutation at >10% VAF (OR 6.5), compared to only ∼4% of controls[37]. Clonal complexity was also greater in pre-AML cases, with ≥1 mutations detectable in nearly 50% of cases compared to only ∼5% of controls ($\gtrsim$1% VAF)[38]. Factors associated with accelerated time to AML progression included clonal complexity (6.9 years vs 9.1 years for ≥ 2 mutations vs 1 mutation) and mutations in specific genes (accelerated time to progression associated with mutations in TP53, DNMT3A, RUNX1 and spliceosome genes)[38].

1.5 Key unanswered questions

These studies captured 'snapshots' of the pre-leukaemic evolutionary process and provided evidence of appreciable clonal differences between individuals who develop AML compared to those who remain cancer-free. Before these findings can be translated into a robust risk-stratification and early detection tool, however, several questions remain unanswered:

How do we explain the wide variation in VAFs observed in clonal haematopoiesis?

Whilst these studies showed that higher VAFs were associated with an increased risk of AML[37,38], simply using VAF measurements to comprehensively risk-stratify specific variants is difficult because the VAF of a specific variant can vary by over 3 orders of magnitude between individuals[7]. What causes this variation? Do the variants confer different growth advantages (fitness effects) in different individuals? Does the bone marrow microenvironment affect the growth of clones? Does the difference in VAF simply reflect the different ages at which the mutations were acquired?

Should all variants within the same gene be considered to confer the same risk?

Pre-AML cases were enriched for mutations in certain genes, relative to controls, suggesting the risk of progressing to AML depends on which gene is mutated[37,38]. However, certain variants within a gene are observed more frequently than others (e.g. R882H in DNMT3A)[7,38] and evidence from mouse models[40,41] and human leukaemia cell lines[42] has shown that the functional consequence of mutations can depend on where in the gene they occur and what effect they have on the transcribed gene product. It is therefore likely that not all variants within a gene confer the same risk of AML. To build robust risk-stratification tools we need to be able to risk stratify individual variants.

What are the key parameters for stem cell evolutionary dynamics?

In order to build quantitative models of AML risk, we need a robust understanding of stem cell dynamics and numbers for key stem cell parameters. What is the population size of HSCs susceptible to acquiring mutations? Is the chance of acquiring a mutation uniform throughout life and, if a mutation occurs, how does it affect stem cell dynamics? Clonal expansion can occur through a bias towards HSC symmetric self-renewal divisions[43–46], but how often do these divisions occur?

HSC mutation rates can be estimated from the number of mutations observed across the genome, with most studies in broad agreement with a genome-wide average HSC mutation rate of $\sim 10^{-9}$ per bp per year[47,48], although variation in mutation rates within genes, across the genome and between individuals will occur[49–51]. Within genes, at the base pair level, the sequence context of the base is felt to be the greatest source of variation in mutation rate, with mutation rates at CpG sites occurring $\sim$10 times more frequently than non-CpG sites[49,52–54] due to spontaneous deamination of 5-methylcytosine. Across the genome, replication timing and chromatin accessibility are felt to play a significant role in mutation rate variability[50,51,55,56]. Between individuals, analysis of the sequence context of mutations has revealed distinct mutational signatures for different types of cancers, some of which reflect particular mutagenic exposures (e.g. smoking, UV light or previous alkylating agent exposure), increasing age or differences in mutation repair efficiency[57–59].

The HSC population size has historically been harder to measure, with most attempts involving extrapolation from animal models or calculation of relative proportions of cells, with estimates ranging from 10^2 to 10^9 cells[60–67].

HSC division rates, particularly symmetric self-renewal division rates, are also hard to measure. Previous attempts have involved extrapolation from mouse estimates[64,65], analysis of the changing X-chromosome inactivation ratios in blood with age[66] and telomere fluorescence measurements[67].

How does sequential accumulation of mutations affect growth rates?

We know that pre-leukaemic mutations appear to be acquired sequentially[17,38,68], but little is known about how different mutation combinations affect growth rates and risk of progression to AML. Mouse and cell-culture models have demonstrated co-operation between different genetic mutations[69–71] and the order in which individuals acquire JAK2 and TET2 mutations has been shown to influence clonal evolution and expansion rates in myeloproliferative disorders[72], but little is known about this in clonal haematopoiesis and AML. To determine how different combinations of mutations interact and affect the evolution of clones over time requires high temporal resolution longitudinal data as well as an understanding of normal stem cell dynamics.

At what age are mutations acquired?

We know that the prevalence of clonal haematopoiesis and AML increases with age[3–5,19], but when does the acquisition of the first mutation occur? Is this different in individuals who develop AML compared to those who remain cancer-free? Retrospective analysis of neonatal Guthrie blood spots from children who had subsequently developed t(8;21) AML between the ages of 3 and 12, found clonotypic genomic AML1-ETO sequences[73]. This provided evidence that pre-leukaemic lesions can occur as early as *in utero*, but is this also the case for adults that develop AML and how long does it take before the second, and subsequent, mutations are acquired?

What is the pattern of pre-leukaemic evolution?

By the time AML is diagnosed, the dominant clone has accumulated ∼5 mutations in recurrently mutated genes[21], but from over 5000 possible driver mutations[20]. What is the pattern of evolution of these mutations? If the HSC population size is large and/or the mutation rate high, we would expect clonal interference to play a key role in early evolution, with multiple lineages growing and competing over time[45]. If the mutations are rare, however, we might expect to see a simple sequential acquisition of mutations over time. We might also expect some pre-leukaemic clones to disappear as a result of immune surveillance[74], but how much evidence is there for this and how does it affect progression to AML?

To try to answer these questions we set out to gain a quantitative understanding of each stage of the step-wise process to AML, from acquisition and clonal expansion of the initial driver mutation (using single timepoint data from ∼50,000-500,000 individuals) all the way through to pre-leukaemic evolution (using longitudinal blood samples from 50 individuals who subsequently developed AML).

In Chapter 2 and Chapter 3 of this thesis, we focused on understanding the acquisition and expansion of the initial clonal haematopoiesis driver mutation, using single nucleotide variant data amassed from ∼50,000 individuals (Chapter 2) and mosaic chromosomal alteration (mCA) data from ∼500,000 individuals (Chapter 3). Using a framework based on a simple branching model of HSC dynamics we found that the wide variation in VAFs observed between individuals can be explained by the combined effects of chance (when a mutation arises) and fitness differences (how fast the mutant clone expands). Most mutations appear to be acquired at roughly a constant rate throughout life and if they confer a positive fitness effect they result in exponential growth. The data, across thousands of individuals, are consistent with the age dependence of clonal haematopoiesis being driven simply by a continuing risk of mutations and subsequent clonal expansions that lead to increased detectability at older ages. Our framework also allowed us to quantify the fitness effects of individual variants, at single nucleotide resolution, the fitness effects of 60% of all possible mCAs and the distribution of fitness effects across genes. We generated a league table of the fittest mutations, which, in support of our inferences, we find confer an increased risk of AML.

To determine how the dynamics of clonal haematopoiesis differs in individuals who progress to AML, we identified longitudinal blood samples that had been collected annually at multiple timepoints from 50 women who subsequently developed AML, as well as age-matched controls who remained cancer free. To study the dynamics of pre-AML mutations over time we needed to not only have the ability to detect a comprehensive array of AML-associated mutations, but also the ability to detect them when they were at very low frequency. In Chapter 4 of this thesis, we describe the development of a custom comprehensive targeted NGS panel, which can detect an array of clonal haematopoiesis and AML-associated genetic changes, including gene mutations, mCAs and chromosomal rearrangements. We used duplex error-corrected sequencing and developed a custom *in silico* noise correction method, which allowed us to call variants down to single molecule resolution. We developed a custom chromosomal rearrangement caller, for accurate translocation and inversion VAF estimation and, by harnessing the power of longitudinal samples, we were able to phase SNPs on an individual basis, allowing us to call mCAs at cell fractions as low as 0.1%.

In Chapter 5 we present our preliminary findings from analysis of this unique set of longitudinal blood samples collected pre-AML diagnosis. We find that there are four main evolutionary patterns in the years preceding diagnosis of AML: linear evolution, evolution with clonal interference, static evolution and late evolution. We calculate the age at acquisition of the first and second mutations and, whilst the initial driver mutation is often acquired early in life, there are some very fit ‘uber drivers’ which appear to occur as the initial event just $\sim$ 4 years pre-diagnosis. We find that the variants we identified as ‘highly fit’ in clonal haematopoiesis (in Chapter 2) are significantly enriched in pre-AML cases and we were able to determine how fitness effects changed with the acquisition of subsequent mutations. These findings reveal key insights into the evolutionary dynamics of clones in the years preceding AML development, which will aid in the development of rational monitoring approaches and identification of individuals who may benefit from early intervention studies.

2

The evolutionary dynamics and fitness landscape of clonal haematopoiesis in healthy individuals

An adapted version of the work presented in this chapter was published in *Science* in 2020.
Watson *et al*. The evolutionary dynamics and fitness landscape of clonal haematopoiesis. *Science* 2020, 367(6485):1449-1454
Analyses undertaken by co-authors are not included here, unless specifically indicated.

2.1 Introduction

The adult haematopoietic system produces more than 2 million blood cells per second, a remarkable feat that is intricately controlled at a hierarchical level to ensure that supply and demand are carefully matched. At the top of the hierarchy is the haematopoietic stem cell (HSC); a long-lived multipotent cell with the ability to self-renew. This long lifespan and capacity for self-renewal means that HSCs are vulnerable to the accumulation of mutations over time, including mutations in cancer-associated genes. Indeed, since a skewed pattern of X-chromosome inactivation was first observed in the blood cells of 20% of healthy elderly women $\sim$ 20 years ago[27], numerous large-scale sequencing studies have shown that the acquisition of leukaemia-associated mutations in our blood cells becomes increasingly common with age[3,4,8,9,75], with deeper sequencing studies showing their presence is almost ubiquitous in those over the age of 65[7,39]. While this phenomenon, 'clonal haematopoiesis', will be of little consequence in most individuals, it has emerged as an important pre-cancerous state[3,4,29], representing the first of a multi-step process that in some individuals can progress to a blood cancer, such as AML. To be able to identify those individuals most at risk we need a better quantitative understanding of this first step of mutation acquisition and clonal expansion and how this affects normal haematopoiesis.

We know that the risk of progressing to blood cancer depends on which gene is mutated[76,77], but should all variants within the same gene be considered equally high risk? We also know that the higher the variant allele frequency (VAF) of a mutation, the greater the risk of progression to AML[76,77], but a challenge to using VAFs to risk-stratify variants is that the VAF of a specific variant can vary by over 3 orders of magnitude between different individuals[7]. How do we explain this variation? Are these differences in VAFs between individuals the result of cell-intrinsic fitness advantages[78], cell-extrinsic perturbations[79] or sheer chance (when a mutation occurs and its stochastic growth after it occurs)[9]? To identify the most highly fit variants we need to understand how mutation, genetic drift, and differences in fitness combine to produce the spectrum of VAFs we observe in clonal haematopoiesis. The advent of higher throughput and more cost-effective sequencing technologies over the past decade has resulted in a significant increase in available blood sequencing data, providing us with the means to try to answer these questions.

Here, using insights from evolutionary biology we present a novel quantitative analysis of VAF spectra using publicly available blood sequencing data amassed from $\sim$50,000 individuals. Determining the growth rate of mutations usually requires multi-timepoint data but, using our framework we show that it is possible to quantify the fitness of key pathogenic variants, down to single-nucleotide resolution, using single-timepoint data. We present a league table of the fittest and potentially most pathogenic variants in blood and quantify the distribution of fitness across key clonal haematopoiesis genes. We reveal the distribution of fitness effects in these genes to be highly skewed, with most mutations conferring either a weak or no fitness effect. Our framework also allows us to make quantitative predictions for the total number of haematopoietic stem cells (HSCs) that maintain the peripheral blood as well as the prevalence of clonal haematopoiesis across ages. While previous studies have suggested the age dependence of clonal haematopoiesis is due to an ageing stem cell niche, we show that the data are consistent with a much simpler explanation: as clones grow, they become easier to detect.

2.2 A branching model of haematopoietic stem cell dynamics

To determine how mutation, drift and selection contribute to the wide variation in VAFs between individuals, we considered a simple stochastic branching model of haematopoietic stem cell (HSC) dynamics based on classic population genetics models[43–46], but adapted to include a spectrum of ages and a spectrum of fitness effects.

In this model, there is a population of N diploid HSCs. In normal homeostasis HSC numbers are maintained, at the population level, by a constant balance of self-renewal and differentiation[43,80,81]. This homeostasis could be achieved if the HSCs divided only asymmetrically (producing 1 HSC and 1 differentiated cell with each division), but this would not be a robust homeostatic mechanism if increased self-renewal was required. Indeed, work over the past decade, predominantly in skin, has shown that stem cells also have the ability to divide symmetrically, producing two differentiated cells or two stem cells with each division[43,82,83]. Therefore in this model the HSCs make 1 of 3 fate decisions with each cell division: symmetrically divide to produce 2 HSCs, asymmetrically divide to produce 1 HSC and 1 differentiated cell or symmetrically divide to produce 2 differentiated cells (Figure 2.1a). The differentiated cells ultimately die and so the average offspring per generation, within the total population, is 1. At the individual cell level the HSC fate decision is random, however, and so there are fluctuations in offspring number at the individual cell level, known as 'drift'[81,84,85]. The overall population HSC number (N) will fluctuate too, but only with a magnitude $\sqrt{1/N}$, which means these fluctuations are suppressed when the total number of HSCs is large. In normal homeostasis, feedback mechanisms likely exist to keep total HSC numbers fairly stable, for example from the stem cell niche[86–88] and extrinsic regulatory molecules[84,89,90].

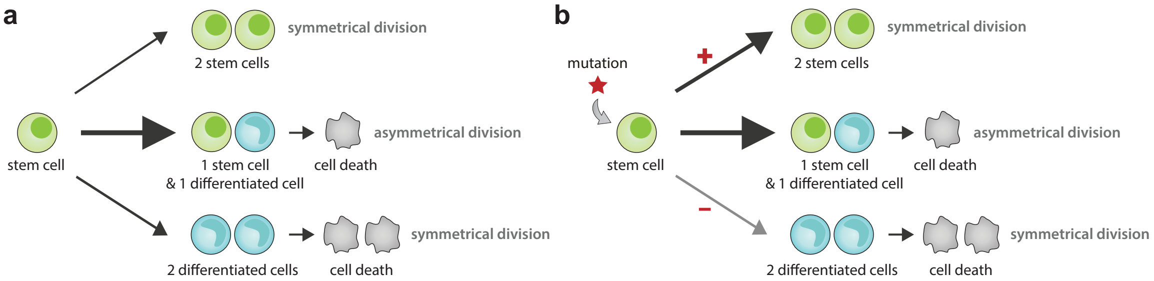


Figure 2.1 Branching model of haematopoietic stem cell dynamics. a. In normal homeostasis the HSC makes 1 of 3 fate decisions with each cell division. **b.** Mutations with a positive fitness effect (red star) cause an imbalance in stochastic cell fates toward self-renewal. This can be an increase in the rate of self-renewal (red plus sign), a decrease in differentiation (red minus sign), or a combination of the two, resulting in clonal expansion.

In this model of HSC dynamics, mutations are acquired at a constant rate (μ), per year. The mutation's fate depends on its influence on the HSC's stochastic fate decisions through a fitness effect (s). Neutral mutations ($s = 0$) do not alter the balance between self-renewal and differentiation, which means that neutral mutations either rapidly go extinct or grow slowly, due to drift, and remain at low VAFs. Beneficial mutations ($s > 0$) increase the rate of self-renewal relative to symmetric differentiation (Figure 2.1b), such that the average offspring per cell per generation increases from 1 to $1+s$. Over time (t years), this causes the lineage to change in size as $(1+s)^t$, a binomial series which $\approx e^{st}$ for small s. Therefore, provided the beneficial mutation escapes stochastic drift to extinction, it will eventually grow exponentially at rate s per year. The relative increase in the rate of self-renewal, caused by beneficial mutations, can be achieved by biasing of the cell fate alone (i.e. increasing the probability of self-renewal) (Figure 2.1b, red plus sign) and/or by decreasing differentiation or apoptosis (Figure 2.1b, red minus sign).

The branching model of haematopoietic stem cell dynamics allows us to infer how large we expect a clone of stem cells to become after a certain period of time and what the expected distribution of clone sizes (VAFs) should be. Variants with a high fitness effect (s) or those acquired early in life are expected to reach high VAFs (trajectories labelled 1 and 2 in Figure 2.2), whereas variants with a low fitness effect or those acquired late in life are restricted to low VAFs (trajectories labelled 3 and 4 in Figure 2.2). This variation in the fitness effect and acquisition age of variants produces a characteristic spectrum of VAFs that can be measured in a single blood sample (Figure 2.2 insets).

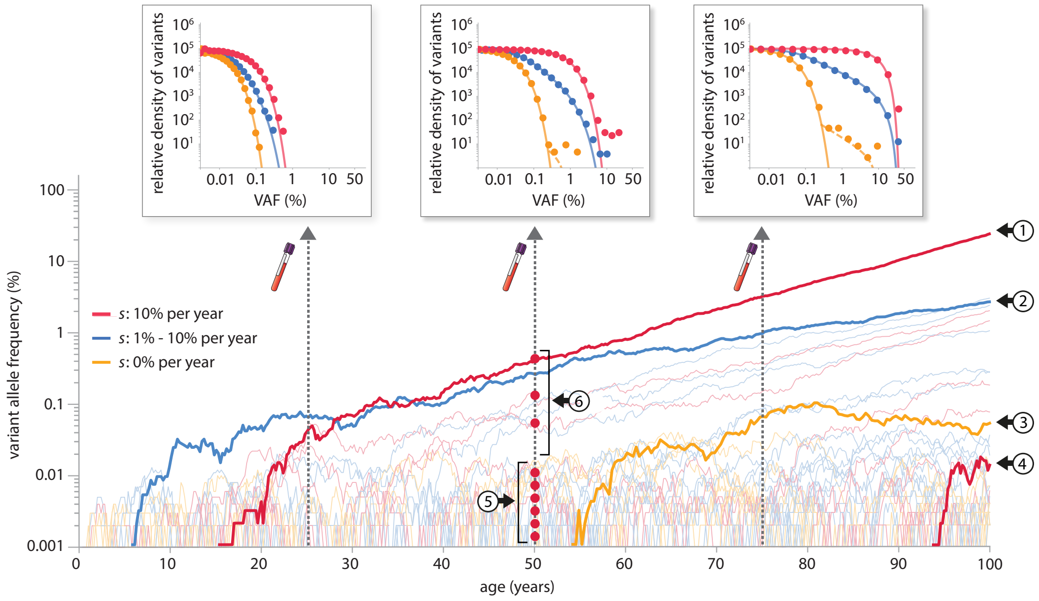


Figure 2.2 The branching model of HSC dynamics produces characteristic variant trajectories and VAF spectra. Simulations of HSC populations, behaving according to a branching model of HSC dynamics, show how differences in fitness effect and age produce characteristic VAF spectra. The vertical dashed lines indicate the timings of the simulated blood samples that produce the VAF spectra shown in the insets. The numbered features are explained in the main text. The red dots labelled 5 and 6 highlight where the red trajectories cross the vertical dashed line.

How these VAF distributions (Figure 2.2 insets) change with age (t), the variant's fitness effect (s) the variant's mutation rate (μ), the population size of HSCs (N) and the time in years between successive symmetric cell differentiation divisions (τ) is given by the following expression for the probability density as a function of $l = \log(VAF)$ (see Appendix A.1):

$$\rho(l) = \frac{\theta}{(1-2e^l)} e^{-\frac{e^l}{\phi(1-2e^l)}} \qquad \text{where } \theta = 2N\tau\mu \quad \text{and} \quad \phi = \frac{e^{st}-1}{2N\tau s} \tag{2.1}$$

$$\text{for neutral mutations (when } s = 0\text{): } \phi = \frac{t}{2N\tau}$$

There are two key features to this distribution. First, because the model assumes the chance of acquiring a specific mutation is uniform throughout life, and beneficial mutations grow exponentially, the trajectories for a specific variant, with the same mutation rate and fitness effect, will be uniformly spaced straight lines when plotted on a log VAF scale (red dots labelled 5 in Figure 2.2). This also means that the density of these variants at a single timepoint is expected to be uniform at low cell fractions, producing a flat density on a log VAF histogram (e.g. red datapoints, Figure 2.2 insets). The number of these low VAF variants (y-intercept) is determined by the product of the mutation rate (μ) and the HSC population size ($N\tau$) (i.e. θ). Dividing the density of variants by the mutation rate (μ) can therefore provide us with an estimate for $N\tau$, and vice versa.

Second, because the age of the oldest surviving variant cannot exceed the age of the individual, there is a characteristic maximum VAF (ϕ) that a specific variant can reach. This is determined by how quickly the affected cells can grow (i.e. the fitness effect, s) and the age of the individual (t). To reach VAFs > ϕ requires a variant to both occur early in life and stochastically drift to high VAFs, which is unlikely. Therefore, the density falls off exponentially for VAFs > ϕ (red dots labelled 6 in Figure 2.2). There is a sharp density fall-off at 50% VAF because even a variant that is present in a very large proportion of total HSCs will tend toward 50% VAF because the cells are diploid.

When considering a number of variants with a range of fitness effects and mutation rates (e.g. all the variants within a specific gene) (Figure 2.2, blue trajectories), the density of variants on a log VAF histogram is still expected to be uniform at low VAF, but the y-intercept will be determined by the product of the *sum* of the variants' mutation rates and the HSC population size ($N\tau$) (Figure 2.2, blue datapoints). The characteristic maximum VAF (ϕ) a variant can reach will be different amongst the variants and so the VAF densities, rather than falling off at a single VAF, will fall-off gradually over a range of VAFs which will be determined by the distribution of fitness effects.

It follows, if we plot the distribution of VAFs observed across large numbers of individuals and compare it to our expected distribution (eq. 2.1), we can determine how consistent HSC behaviour is with our simple model. If consistent, fitting our expected distribution (eq. 2.1) to the data will enable us to infer HSC numbers and division times ($N\tau$), mutation rates (μ) and fitness effects (s) for specific variants as well as the distribution of fitness effects within genes.

2.3 The spectrum of VAFs across ∼50,000 individuals

To infer these parameters and test the predictions of our model we used peripheral blood sequencing data from ∼ 50,000 individuals, amassed from nine publicly available blood sequencing datasets[3,4,6–9,38,39,75] (Table 2.1). VAF measurements in bone marrow and peripheral blood show good concordance[91] and so peripheral blood VAF measurements were used as a proxy to reflect clonal composition at the level of the bone marrow HSCs.

The nine studies we included in our analysis varied in their number of participants and sequencing depth. Most large-scale studies were limited by standard sequencing error rates and were only able to detect VAFs $> 3\%$[3,4] while smaller studies, using error-correcting techniques, were able to detect VAFs as low as 0.03%[7,8,39] (Figure 2.3).

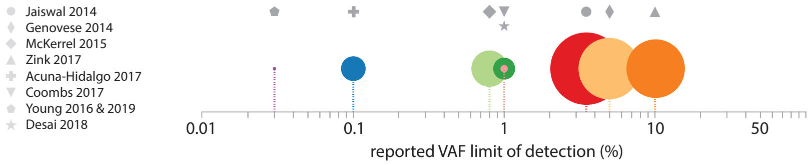


Figure 2.3 Study sizes and VAF limits of detection. Studies used in this analysis varied in the number of participants (indicated by relative circle size) and reported VAF detection thresholds.

As predicted, the spectrum of VAFs across all individuals was broad, even within the same gene, as shown for DNMT3A, the most commonly mutated clonal haematopoiesis gene (Figure 2.4). The majority of variants were found at low VAF and some variants were observed far more frequently than others, for example DNMT3A R882H (Figure 2.4, red datapoints). Most observed variants were nonsynonymous or frameshift variants, with synonymous variants being rare. Some of the paucity of synonymous variants could be attributable to most studies excluding them. However, in the three studies that did report them[7,8,39], they tended to be restricted to low VAFs.

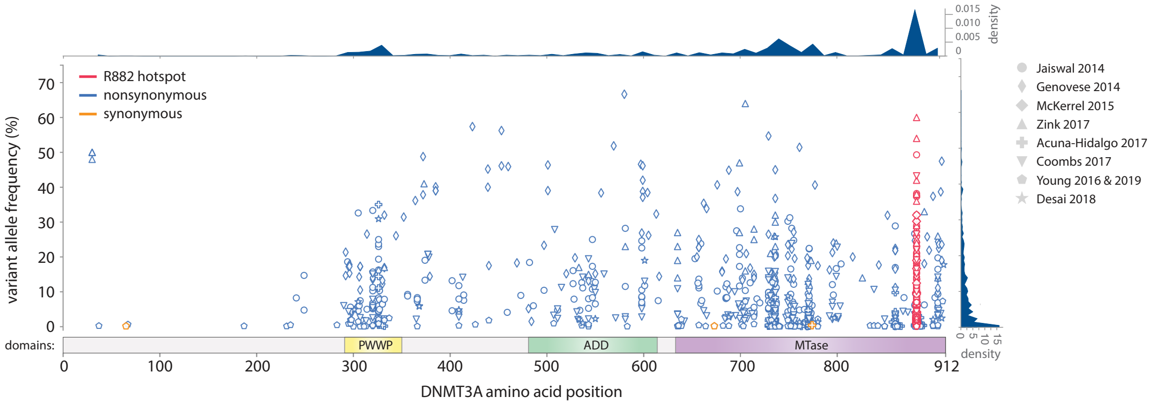


Figure 2.4 Distribution of VAFs across DNMT3A from blood sequencing data from ∼50,000 individuals. The majority of variants are found at low VAF, but the spectrum of VAFs varies by >3 orders of magnitude across individuals.

Table 2.1 Details of studies included in our analysis. ECS: Error-corrected sequencing, HPFS: Health Professionals Follow-Up Study, MSKCC: Memorial Sloan Kettering Cancer Center, NBS: Nijmegen Biomedical Study, NHS: Nurses Health Study, smMIPs: Single Molecule Molecular Inversion Probes, T2DM: Type 2 Diabetes Mellitus, UKHLS: UK Household Longitudinal Study, WES: Whole Exome Sequencing, WGS: Whole Genome Sequencing, WHI: Women's Health Initiative, WTCCC: Wellcome Trust Case Control Consortium

Study	Study size	Participants	hematological characteristics	Age range	Sex	Sequencing method	Reported limit of SNV detection	Participants included in our data analysis
Jaiswal 2014[3]	17182	Population-based cohorts (15801 cases/controls from T2DM association studies, 1381 from Jackson Heart study)	Unselected for hematologic phenotype.	19-108	51% female, 49% male	WES	3.5%	All (17182)
Genovese 2014[4]	12380	Swedish individuals (6245 controls, 4970 schizophrenia, 1165 bipolar disorder)	Unselected for hematologic phenotype.	19-93	49% female, 51% male	WES	5%	All (12380)
McKerrel 2015[6]	4219	3067 UK blood donors (WTCCC), 1152 UKHLS participants	Unselected for hematologic phenotype.	17-98	?	Barcoded multiplex PCR of mutational hotspots.	0.8%	All (4219)
Young 2016[7]	20	Population-based cohort (NHS) - all had blood sample at 2 time-points	Unselected for hematologic phenotype, but no history of cancer or major chronic disease.	51-74	100% female	ECS (adapted Illumina TruSight Myeloid Panel).	0.03%	All (only 1 time-point)
Zink 2017[9]	11262	Icelanders participating in various disease projects at deCODE genetics.	Excluded from study if diagnosis of hematological malignancy before or within 6 months after blood sample.	<10 - >100	55% female, 45% male	WGS and then targeted resequencing on Illumina TruSight Myeloid Panel of 72 individuals with CH.	10% (WGS)	All (11262) (only WGS)
Acuna-Hidalgo 2017[8]	2006	Population-based cohort (NBS).	Unselected for hematologic phenotype.	20-69	50% female, 50% male	smMIPs	0.1%	All (2006)
Coombs 2017[75]	5649	Patients with non-hematologic cancers at MSKCC.	Excluded from study if active hematologic cancer or precursor conditions.	<1-98	51% female, 49% male	Hybridization capture-based sequencing assay encompassing all protein-coding exons of 341 or 410 cancer-associated genes (MSK-IMPACT).	1%	Only those who were chemotherapy-naïve and radiotherapy-naïve (1591)
Desai 2018[38]	424	Individuals who developed AML and controls (~1:1 ratio) within WHI cohort - >200 with samples from $\geq$2 time-point.	Controls excluded if history of hematologic disorder or AML during study.	51-79	100% female	Targeted sequencing using custom capture probes (Nimblegen).	1%	Only controls (181) - only 1 time-point (baseline sample)
Young 2019[39]	103	Individuals who developed AML and controls (2:1 ratio) within NHS and HPFS cohorts - all NHS participants had sample from 2 time-points.	Controls excluded if history of cancer (other than non-melanoma skin cancer).	48-79	45% female, 55% male	ECS (adapted Illumina TruSight Myeloid Panel)	0.03%	Only controls (69) (only 1 time-point included)
						TOTAL INCLUDED IN OUR ANALYSIS:		48910

2.3.1 Combining data from nine different studies

To test the predictions of our model we needed to plot VAF distributions for SNVs from all the studies together. The studies not only varied in their number of participants but also in their panel footprint, with some studies choosing to target only hotspot sites[3,6], some focusing on specific exons and others including entire coding regions (Appendix A.2). To meaningfully combine the data from all the studies, we had to control for these differences. To do this we normalised the number of observed variants by the size of the study, as well as the total study-specific mutation rate (for variant or gene of interest), controlling for the trinucleotide contexts of mutations.

Mutation rate estimates

To estimate variant-specific and study-specific mutation rates, we used recently published whole-genome sequencing data from 140 single-cell derived HSC colonies derived from a healthy 59 year old man[48]. A total of 129,582 genome-wide somatic mutations were observed across the 140 colonies which, over 59 years of life, equates to $\approx$ 15.7 mutations per year per cell ($\approx 2.7 \times 10^{-9}$ per bp per year). The observed substitutions were categorised into 96 trinucleotide-context specific categories according to the pyrimidine base change and its neighbouring 5' and 3' bases (e.g. A[C>A]A). To obtain site-specific mutation rate estimates (per year), for these 96 site contexts, as well as their complementary site contexts, we normalised by the trinucleotide frequencies (of both sites) across the mappable genome (5.87×10^9 bp per cell) (Appendix A.3.1). The normalised number of substitutions was then divided by the number of colonies (140) and the age of the individual (59), in order to obtain a haploid trinucleotide-context-site-specific mutation rate in units of years:

$$\text{site-specific (e.g. A[C>A]A) mutation rate} = \frac{\text{observed number of substitutions}}{(\text{trinuc. freq} + \text{complementary trinuc. freq}) \times (5.87 \times 10^9) \times 140 \times 59}$$

Having calculated mutation rates for all possible trinucleotide-context specific base changes (Table 2.2), the mutation rate for individual variants could be estimated, for example the mutation rate of DNMT3A R882H, whose trinucleotide-context base-change is C[G>A]C, is 18.82×10^{-9} per year (Appendix A.3.2). To determine study-specific mutation rates across a particular gene, e.g. TET2, the regions targeted by the study were determined from the study's published information and then all the site-specific mutation rates in these regions were summed (Appendix A.3).

We only included single nucleotide variants (SNVs) in our analysis, due to mutation rate uncertainties for other classes of mutation.

Table 2.2 Site-specific haploid mutation rates according to trinucleotide context of base change. Although we calculate the haploid mutation rates for all 192 site contexts, the rates at particular sites and their complementary partner (e.g. A[C>A]A and T[G>T]T) cannot be distinguished since only the sum of their rates is measured. However this is the relevant rate for calculating how frequently a site mutates since either strand could have undergone the mutation.

Site	**Site-specific mutation rate (μ) ($\times10^{-9}$ /year)**	**Site**	**Site-specific mutation rate (μ) ($\times10^{-9}$ /year)**
A[C>A]A or T[G>T]T	1.33	**A[T>A]A or T[A>T]T**	0.43
A[C>A]C or G[G>T]T	0.95	**A[T>A]C or G[A>T]T**	0.81
A[C>A]G or C[G>T]T	1.06	**A[T>A]G or C[A>T]T**	0.50
A[C>A]T or A[G>T]T	0.76	**A[T>A]T or A[A>T]T**	0.37
C[C>A]A or T[G>T]G	1.54	**C[T>A]A or T[A>T]G**	0.42
C[C>A]C or G[G>T]G	0.81	**C[T>A]C or G[A>T]G**	0.38
C[C>A]G or C[G>T]G	1.16	**C[T>A]G or C[A>T]G**	0.45
C[C>A]T or A[G>T]G	0.98	**C[T>A]T or A[A>T]G**	0.38
G[C>A]A or T[G>T]C	1.12	**G[T>A]A or T[A>T]C**	0.41
G[C>A]C or G[G>T]C	0.64	**G[T>A]C or G[A>T]C**	0.45
G[C>A]G or C[G>T]C	0.85	**G[T>A]G or C[A>T]C**	0.31
G[C>A]T or A[G>T]C	0.77	**G[T>A]T or A[A>T]C**	0.40
T[C>A]A or T[G>T]A	0.54	**T[T>A]A or T[A>T]A**	0.29
T[C>A]C or G[G>T]A	0.54	**T[T>A]C or G[A>T]A**	0.27
T[C>A]G or C[G>T]A	0.91	**T[T>A]G or C[A>T]A**	0.22
T[C>A]T or A[G>T]A	0.50	**T[T>A]T or A[A>T]A**	0.25
A[C>G]A or T[G>C]T	1.09	**A[T>C]A or T[A>G]T**	1.47
A[C>G]C or G[G>C]T	0.48	**A[T>C]C or G[A>G]T**	0.83
A[C>G]G or C[G>C]T	0.96	**A[T>C]G or C[A>G]T**	1.49
A[C>G]T or A[G>C]T	0.73	**A[T>C]T or A[A>G]T**	1.35
C[C>G]A or T[G>C]G	0.46	**C[T>C]A or T[A>G]G**	0.88
C[C>G]C or G[G>C]G	0.46	**C[T>C]C or G[A>G]G**	0.81
C[C>G]G or C[G>C]G	0.66	**C[T>C]G or C[A>G]G**	0.81
C[C>G]T or A[G>C]G	0.44	**C[T>C]T or A[A>G]G**	0.74
G[C>G]A or T[G>C]C	0.54	**G[T>C]A or T[A>G]C**	0.88
G[C>G]C or G[G>C]C	0.40	**G[T>C]C or G[A>G]C**	0.89
G[C>G]G or C[G>C]C	0.63	**G[T>C]G or C[A>G]C**	0.65
G[C>G]T or A[G>C]C	0.44	**G[T>C]T or A[A>G]C**	1.09
T[C>G]A or T[G>C]A	0.28	**T[T>C]A or T[A>G]A**	0.62
T[C>G]C or G[G>C]A	0.44	**T[T>C]C or G[A>G]A**	0.54
T[C>G]G or C[G>C]A	0.57	**T[T>C]G or C[A>G]A**	0.54
T[C>G]T or A[G>C]A	0.47	**T[T>C]T or A[A>G]A**	0.50
A[C>T]A or T[G>A]T	3.21	**A[T>G]A or T[A>C]T**	0.20
A[C>T]C or G[G>A]T	3.18	**A[T>G]C or G[A>C]T**	0.15
A[C>T]G or C[G>A]T	32.69	**A[T>G]G or C[A>C]T**	0.36
A[C>T]T or T[G>A]T	3.10	**A[T>G]T or A[A>C]T**	0.19
C[C>T]A or T[G>A]G	1.96	**C[T>G]A or T[A>C]G**	0.17
C[C>T]C or G[G>A]G	2.65	**C[T>G]C or G[A>C]G**	0.19
C[C>T]G or C[G>A]G	14.15	**C[T>G]G or C[A>C]G**	0.30
C[C>T]T or A[G>A]G	4.38	**C[T>G]T or A[A>C]G**	0.25
G[C>T]A or T[G>A]C	2.32	**G[T>G]A or T[A>C]C**	0.19
G[C>T]C or G[G>A]C	3.45	**G[T>G]C or G[A>C]C**	0.12
G[C>T]G or C[G>A]C	18.82	**G[T>G]G or C[A>C]C**	0.25
G[C>T]T or A[G>A]C	4.09	**G[T>G]T or A[A>C]C**	0.17
T[C>T]A or T[G>A]A	0.99	**T[T>G]A or T[A>C]A**	0.16
T[C>T]C or G[G>A]A	1.56	**T[T>G]C or G[A>C]A**	0.16
T[C>T]G or C[G>A]A	12.00	**T[T>G]G or C[A>C]A**	0.28
T[C>T]T or A[G>A]A	1.40	**T[T>G]T or A[A>C]A**	0.19

Data trimming below study-specific limits of detection

Whilst studies generally reported their VAF detection threshold, this is typically determined by a predetermined false positive rate, at which false negative rates could be substantial. Because an accurate VAF density measurement, particularly at low VAF, is important for the fitting of our theory distribution, it was important for us to only include variants at VAFs where the false negative rate was expected to be low. To estimate where false negative rates were beginning to have a substantial effect on the data, we used variants in DNMT3A (which had the most data) and chose a threshold VAF below which the density began to decline (Figure 2.5). Variants were excluded from our analysis if their VAF was below this study-specific VAF threshold and the same threshold was used for trimming all other variants reported by the study (Table 2.3).

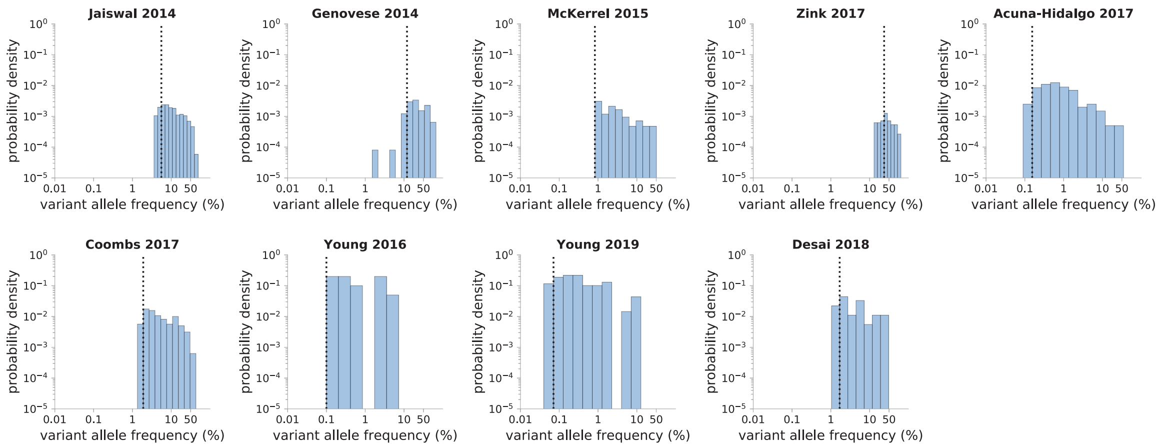


Figure 2.5 Data trimming of DNMT3A variants . Vertical dashed lines on the probability density histograms indicate the VAF below which the density of DNMT3A variants starts to fall off and thus likely represents the study's limit of reliable variant detection. DNMT3A variants at VAFs lower than this cut-off were not included in our data analysis and this cut-off was also used for trimming all other variants reported by that study.

Table 2.3 Study-specific limits of detection used for data trimming. The VAF limit of SNV detection reported by each study is shown in comparison to the study-specific threshold below which we trimmed each study's data.

Study	Reported VAF limit of SNV detection	VAF threshold used for data trimming
Jaiswal 2014[3]	3.50%	5.56%
Genovese 2014[4]	5.00%	11.94%
McKerrel 2015[6]	0.80%	0.83%
Zink 2017[9]	10.00%	23.63%
Coombs 2017[75]	1.00%	1.90%
Acuna-Hidalgo 2017[8]	0.10%	0.16%
Young 2016[7]	0.03%	0.10%
Desai 2018[38]	1.00%	1.69%
Young 2019[39]	0.03%	0.07%

2.4 Haematopoietic stem cell numbers and division times

Testing the model and inferring HSC numbers using DNMT3A R882H

To infer HSC numbers and test the predictions of our model, we first focused on a single variant, DNMT3A R882H, which was the most commonly observed variant across all studies, being detected in 105 individuals. Because this single variant is expected to confer the same fitness effect in all individuals, it served as a useful simple first check on our model.

A probability density histogram was plotted, as a function of log VAF (Figure 2.6a, datapoints). To enable the data from all studies to be combined, the densities were normalised by dividing by [number of individuals in the study × bin widths] and, in order to read $N\tau$ from the y-intercept, the densities were rescaled by dividing by 2μ, where μ is the trinucleotide-context-site-specific mutation rate for DNMT3A R882H (Appendix A.3.2). Estimates for $N\tau$ and s were inferred using a maximum likelihood approach (Figure 2.6b-c), minimising the L2 norm between the log rescaled densities and predicted densities, for all datapoints. To account for the distribution of ages across all the studies, the predicted density was calculated by integrating the theoretical density (eq. 2.1) across the distribution of ages, normally distributed with mean 55 years and standard deviation σ. Because the standard deviation of ages (σ) across the studies was unknown, σ was optimised along with $N\tau$ and s.

Consistent with the predictions of our model (Figure 2.2, insets), the density of DNMT3A R882H variants is seen to be flat over almost the entire VAF range (<15% VAF) (Figure 2.6a) and then falls off exponentially. The inferred y-intercept of $N\tau$ is $\approx 100,000 \pm 35,000$ years which, encouragingly, is in close agreement with the number recently inferred from two different single-cell HSC phylogeny studies[48,92].

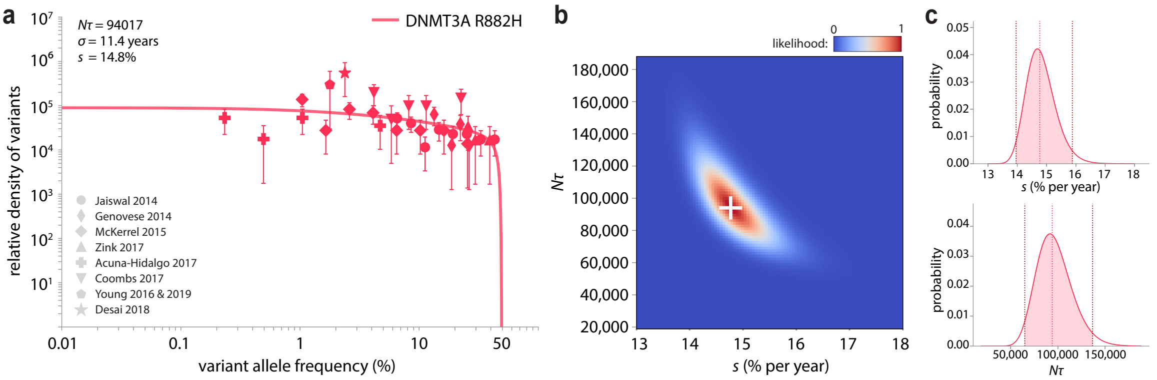


Figure 2.6 Parameter estimation for DNMT3A R882H. **a.** Probability density histogram for R882H with theory distribution (eq. 2.1) fitted using maximum likelihood estimates. The mean age was fixed at 55 years (normally distributed) and maximum likelihood approaches were used to infer the standard deviation of ages (σ) = 11.4 years, $N\tau$ = 94017 and s = 14.8% per year. **b.** Maximum likelihood heatmap for $N\tau$ and s estimates for DNMT3A R882H. White cross marks the most likely $N\tau$ (94017) and s (14.8% per year). **c.** Distribution of likelihoods for s and $N\tau$. Red vertical line represents most likely value. 95% confidence intervals are shown shaded in pink: $N\tau$ 64956 – 136751, s 14.0 – 15.9%.

An important point to note is that ours and other population genetic analyses can only reliably infer the combination $N\tau$ (number of HSCs multiplied by the average time between HSC self-renewing divisions in years) and not N or τ separately. Estimating τ independently of N is challenging. Early developmental mutations suggest that HSCs acquire $\approx$1.2 mutations per cell division[48], which, combined with an HSC mutation rate in adulthood of $\approx$16 mutations per cell per year[48] suggests HSCs divide $\approx$13 times per year. Although this includes both symmetric and asymmetric divisions, it provides us with a lower bound for the time between symmetric self-renewing divisions ($\tau \geq 1/13$ years) and thus an upper-bound on the number of HSCs (N). If $N\tau \approx$100,000, this suggests that <1.3 million HSCs maintain the peripheral blood. A lower-bound for the number of HSCs can be estimated using the largest fitness effect observed in the data. If the HSC fate decisions were completely biased towards self-renewal, which occurred every τ years, then the growth rate would be $s = 1/\tau$. This means that τ cannot be >1/s. Our maximum inferred s was $\approx$ 25% (Figure 2.8) suggesting that $\tau < 4$ years. If $N\tau \approx$100,000 this suggests a lower bound of 25,000 for the number of HSCs.

Validating the model and HSC numbers using synonymous variants

To validate our estimates for $N\tau$, we turned to the VAF distribution of synonymous variants. Because synonymous variants are expected to be functionally neutral ($s = 0$), their characteristic maximum VAF (ϕ) is expected to increase only linearly with age (t), as it is driven by drift alone (see Figure 2.2, insets) and $N\tau$ is the time it would take for a neutral mutation to drift to fixation by chance:

$$\phi = \frac{e^{st}-1}{2N\tau s} \quad \Rightarrow \quad \frac{t}{2N\tau} \quad \text{when} \quad s = 0 \tag{2.2}$$

If haematopoietic stem cells behave according to our model, and our inferred value of $N\tau \approx 100,000$ years is correct, then we would expect the majority of synonymous variants to be found at VAFs less than $\phi = t/2N\tau \approx 0.025\%$ at age 50.

To check whether this was the case, we looked at the log VAF probability density histogram of all synonymous variants reported in the three studies that reported them[7,8,39] (Figure 2.7). The VAF densities for each study were normalised by the study size and, to take in to account the different study panel sizes (which would affect the number of synonymous variants detected), the densities were also normalised by the estimated study-specific synonymous mutation rate (Appendix A.3). An estimate for ϕ was inferred using a maximum likelihood approach, fixing $N\tau$ to that inferred from DNMT3A R882H ($N\tau \approx$ 100,000) and minimising the L2 norm between the log rescaled densities and the predicted densities, taking in to account the distribution of ages across the three studies.

If all synonymous variants were included (Figure 2.7a), the inferred value of ϕ was 0.12%, which was 4.8-fold higher than the 0.025% predicted by our model. However, the distribution of synonymous variants >0.25% VAF are consistent with variants that have either hitchhiked to high VAFs with an expanding fit clone, were acquired early in development (Figure 2.18 orange dashed line, work by Gladys Poon in the Blundell Lab[93,94]), have a functional consequence themselves (e.g. due to codon usage bias) or are in fact nonsynonymous in an alternatively spliced transcript. Because these synonymous variants are not expected to behave neutrally, we excluded them from the maximum likelihood analysis. When only synonymous variants <0.25% VAF were included (Figure 2.7b), the maximum likelihood inferred ϕ was $\approx 0.03 \pm 0.005\%$, which broadly agrees with our model predictions. This internal consistency check indicates that both synonymous and DNMT3A R882H variants point toward similar values of $N\tau$.

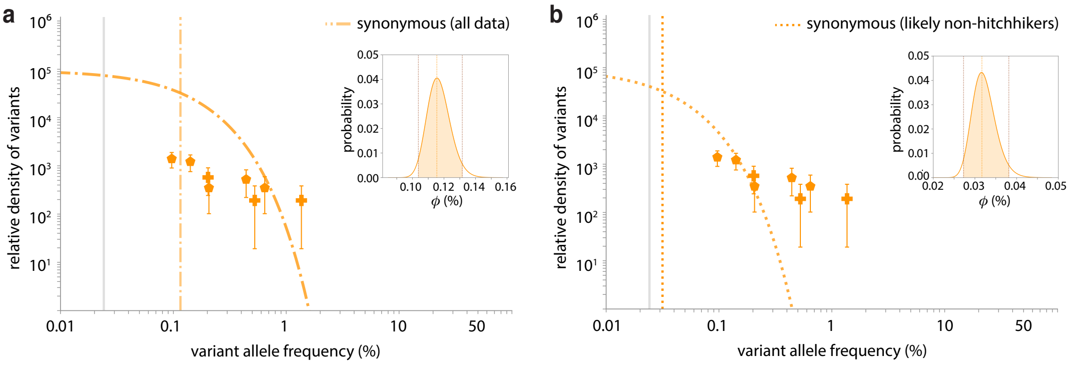


Figure 2.7 Fitted theory distribution for synonymous variants. **a.** If synonymous variants of any VAF are included in the maximum likelihood approach, the inferred ϕ value (orange dot-dashed line) is 0.115% (95% C.I. 0.104-0.132%, inset orange shaded area), which is 4.8-fold higher than the predicted ϕ value of 0.024% (grey vertical line). **b.** If synonymous variants >0.25% VAF are assumed to be hitchhikers, maximum likelihood approaches on those <0.25% VAF infer a ϕ value (orange dashed line) of 0.032% (95% C.I. 0.027-0.038%, inset orange shaded area), which is only 1.3-fold higher than the predicted ϕ value of 0.024% (grey vertical line). Each study is represented by a shaped symbol as described in Fig. 2.6.

2.5 The fitness landscape of clonal haematopoiesis

2.5.1 Fitness effects of single nucleotide variants

Because our model assumes a variant's mutation rate is uniform throughout life, the density of variants is expected to be flat at low VAFs and then to fall off exponentially at a characteristic maximum VAF (ϕ), which is determined by the variant's fitness effect (s). By estimating ϕ from a variant's spectrum of VAFs, we can therefore infer a variant's fitness. This can be illustrated using DNMT3A R882H as an example (Figure 2.6). In agreement with our model, the density of DNMT3A R882H variants is flat at low VAFs and then begins to fall off exponentially at VAFs >12%. Using a maximum likelihood approach (described in Section 2.4), we can estimate that this is consistent with a fitness effect of $\approx 15 \pm 1\%$ per year.

To reveal the fitness landscape of other highly fit and possibly pathogenic variants, we applied this analysis to each of the 20 most commonly observed variants across all studies (Figure 2.8a). Because our estimate for $N\tau$ agrees with other independent estimates[48,92], we fixed $N\tau$ to $\approx$ 100,000 and used a maximum likelihood approach to infer variant-specific s and μ values, minimising the L2 norm between the log rescaled densities and the predicted densities and taking into account the distribution of ages across studies (Appendix A.4.1). Of note, for most variants, the inferred value of μ agreed to within a factor of 5 to that estimated by the site-specific trinucleotide context (Figure 2.8b).

We found that variants in the spliceosome genes SF3B1 and SRSF2 are some of the fittest in clonal haematopoiesis, with fitness effects as high as $s \approx 23\%$ per year, but they are relatively rare due to low mutation rates of $\sim 3 \times 10^{-10}$ per year. Our analysis also reveals that DNMT3A R882H is the most commonly observed clonal haematopoiesis variant, not because it is the most fit, but because it is both highly fit and has a high mutation rate, most likely due to its CpG context (C[G>A]C). The potential of our analyses is highlighted by the GNB1 K57E variant. While this variant has received little attention in clonal haematopoiesis, we find it is in fact highly fit and, importantly, strongly associated with myeloid cancers and potentially targetable[95].

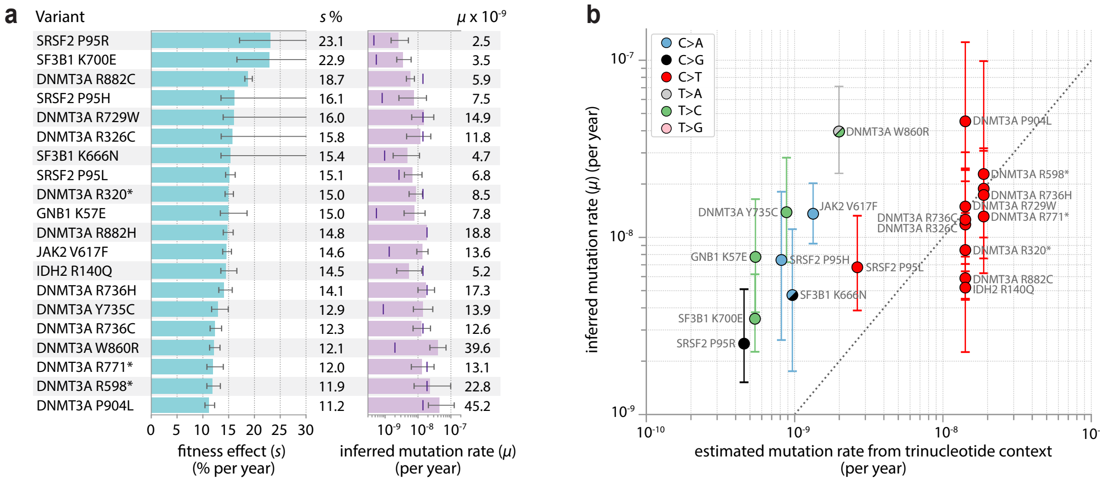


Figure 2.8 The fitness landscape of clonal haematopoiesis variants. **a.** Inferred fitness effects and mutation rates for the top 20 most commonly observed clonal haematopoiesis variants. Error bars represent 95% confidence intervals. Purple vertical lines indicate site-specific mutation rates inferred from trinucleotide context (Appendix A.3). **b.** Inferred mutation rates compared to mutation rates estimated from the variant's trinucleotide context (Table 2.2).

Disentangling the relative effects of mutation rate and fitness effect

A key feature of our framework is its ability to disentangle the relative effects of mutation rate and fitness effect. For example, DNMT3A R882H, which was observed in 105 individuals across the nine studies, is the most commonly observed variant in DNMT3A, followed by R882C, which was observed in 61 individuals. The estimated mutation rate for DNMT3A R882H, from the trinucleotide-context of its base change, is 1.9×10^{-8}, compared to 5.9×10^{-9} for DNMT3A R882C (Appendix A.2). But is R882H's higher prevalence explainable by a higher mutation rate, a higher fitness effect or both? Plotting the distribution of VAFs for each of these two variants and using maximum likelihood approaches, as described above, to infer fitness effects, reveals that R882C actually has a higher fitness effect than R882H (Figure 2.9). So, although R882H variants are observed most frequently, this is attributable to a high mutation rate in combination with a high fitness whereas R882C mutations actually have a higher fitness effect and are thus potentially more pathogenic.

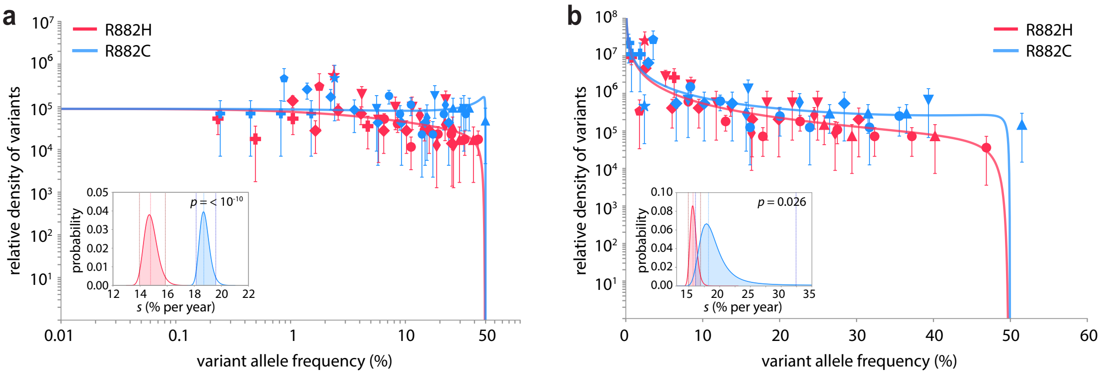


Figure 2.9 Fitted theory distributions for DNMT3A R882H and R882C. **a.** Probability density log-log histograms for R882H (red data) and R882C (blue data), as a function of log VAFs. $N\tau$ was fixed to $\approx$ 100,000 and maximum likelihood approaches were used to infer s of 14.8% (95%C.I. 14.0-15.9%, pink shaded area) for R882H and s of 18.7% (95% C.I.18.1-19.6%, blue shaded area) for R882C. **b.** Probability density linear-log histograms for R882H (red data) and R882C (blue data), as a function of linear VAFs. $N\tau$ was fixed to $\approx$ 100,000 and maximum likelihood approaches were used to infer s of 15.7% (95% C.I. 14.9-16.9 %, pink shaded area) for R882H and s of 18.2% (95% C.I. 16.1-32.4%, blue shaded area) for R882C. Each study is represented by a shaped symbol as described in Fig. 2.6. p-values were calculated from the area under the distribution of difference probability curve where the difference ≤ 0.

2.5.2 Distribution of fitness effects across key clonal haematopoiesis driver genes

To reveal the overall fitness landscapes of key clonal haematopoiesis driver genes, we considered the VAF distribution of all nonsynonymous variants in each of the genes DNMT3A, TET2, ASXL1 and TP53 (Appendix A.4.2). For DNMT3A, the density of nonsynonymous variants at low VAF is broadly consistent with the same $N\tau \approx 100,000$ years inferred from DNMT3A R882H variants (Figure 2.10a). With increasing VAF, however, the density of variants gradually declines, consistent with the predictions of our model, where the density of variants declines over a range of ϕ due to the set of variants having a range of fitness effects (Figure 2.2 insets, blue data).

To estimate the spectrum of fitness effects, log VAF probability density histograms were plotted for all the nonsynonymous variants observed in each study within a particular gene (e.g. DNMT3A) and normalised by the study size. To account for different studies targeting different amounts of the gene of interest, the densities of variants within each study were also normalised by the study-specific gene mutation rate (Appendix A.3). $N\tau$ was fixed to that inferred from DNTM3A R882H ($N\tau \approx$ 100,000) and the distribution of fitness effects was parameterised using a family of stretched exponential distributions with a maximum $s = s_{\text{max}}$:

$$\rho(s) \sim \exp\left[-\left(\frac{s}{d}\right)^{\beta}\right] \tag{2.3}$$

A maximum likelihood approach was then used to optimise the shape (β) and scale (δ) of the distribution, as well as s_{max}.

We found that the spectrum of fitness effects for nonsynonymous variants in DNMT3A was very broad with $\approx$ 40% of variants conferring moderate to high fitness effects ($s > 4\%$ per year, Figure 2.10b).

In contrast, the genes TET2, ASXL1 and TP53 have a spectrum that is more skewed towards low fitness effects with only ≈ 7–10% of all possible nonsynonymous variants in these genes conferring moderate or high fitness effects. These distributions highlight that, in these clonal haematopoiesis genes, most nonsynonymous variants have a low enough fitness that they are effectively neutral, while an important minority have fitness effects great enough to allow them to expand to overwhelm the bone marrow over a human lifespan.

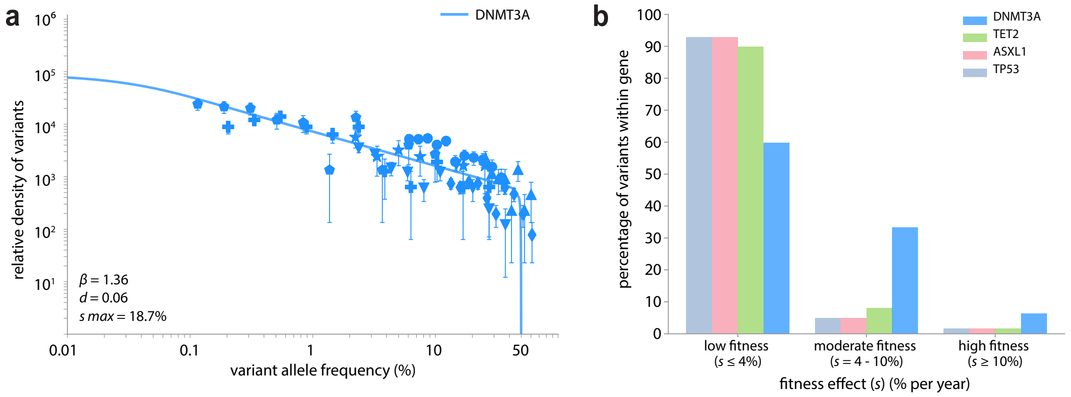


Figure 2.10 The fitness landscape of clonal haematopoiesis genes. a. b. The distribution of fitness effects of non-synonymous variants in key clonal haematopoiesis driver genes, inferred by fitting a stretched exponential distribution (Appendix A.4.2) and dividing this into three fitness classes (low, moderate and high). Over a human lifespan, variants with fitness effects <4% expand only a modest factor more than a neutral variant (low fitness), variants with fitness effects of 4-10% per year expand by substantial factors (moderate fitness), and variants with fitness effects >10% per year can expand enough to overwhelm the bone marrow (high fitness).

2.5.3 Fitness effects within TP53 domains

A recent study found that TP53 missense variants in the DNA binding domain exerted a strong dominant negative effect, leading to clonal expansion of haematopooetic stem and progenitor cells (HSPCs) in mice, whereas missense variants outside the DNA-binding domain had either a neutral or loss-of-function effect[42]. It stands to reason, therefore, that missense variants in the DNA binding domain might have higher fitness effects compared to missense variants outside the DNA-binding domain. Our framework allows us to test this. We considered all TP53 variants in the seven studies that reported TP53 variants[3,4,7–9,39,75] and tested whether the number of variants observed within the DNA-binding domain was significantly different to the number observed outside the DNA binding domain, controlling for the differences in panel coverage and estimated mutation rates in these regions. Because of the differences in sensitivities of each of the studies that reported TP53 variants, the p-values were calculated independently and then combined using Fisher's method (Figure 2.11). The origin of the non-significant p-values in Jaiswal 2014[3], Genovese 2014[4], Zink 2017[9] and Acuna-Hidalgo 2017[8] is the small number of sites outside the DNA-binding domain that were targeted by these studies (Appendix A.2). Taken together, however, these studies demonstrate a significant and substantial enrichment for missense variants in the DNA binding domain ($p = 0.0002$, Poisson exact test followed by Fisher's method), consistent with these variants having higher fitness effects.

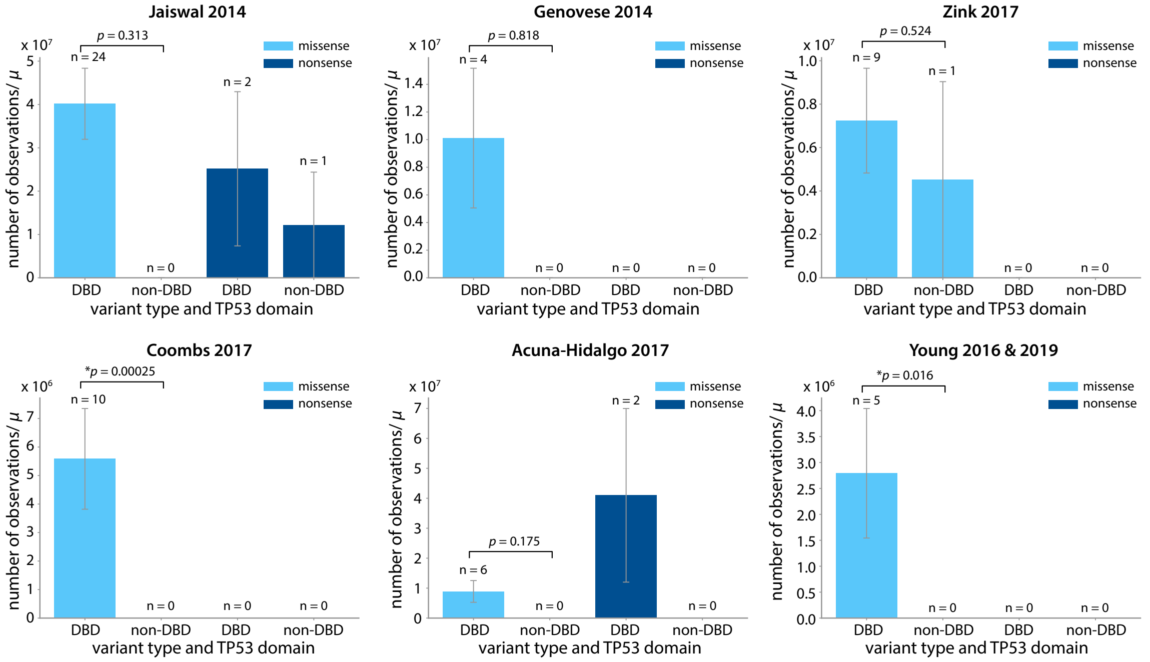


Figure 2.11 Relative frequency of missense and nonsense TP53 variants inside and outside the DNA-binding domain. The relative frequency controls for differences in mutation rate (μ) inside the DNA-binding domain and outside the DNA-binding domain (calculated by summing site-specific mutation rates for the sites included in each study). The actual number of observations in each study is shown above each bar. Error bars are one standard deviation assuming Poisson sampling noise. p-values calculated using the Poisson Exact Test. Combining these p-values using Fisher's method demonstrates a significant and substantial enrichment for missense variants in the DNA binding domain ($p = 0.0002$, Poisson exact test followed by Fisher's method). DBD: DNA-binding domain, non-DBD: non-DNA-binding domain.

2.6 Highly fit variants confer an increased risk of AML

One of the principles underlying pre-cancerous mutation acquisition and clonal expansion is that the greater the fitness effect of a mutation, the faster the clone will expand and the more likely it is that subsequent mutations will be acquired within the same clone. We therefore wondered whether high fitness clonal haematopoiesis variants conferred an increased risk of AML development. By considering the pre-AML and control samples from three studies[38,39,77] we found that individuals harbouring one or more of the 20 highly fit variants we identified (Figure 2.8) are $\approx$ 4-fold more likely to develop AML compared to those harbouring lower fitness variants ($p < 10^{-5}$) (Figure 2.12).

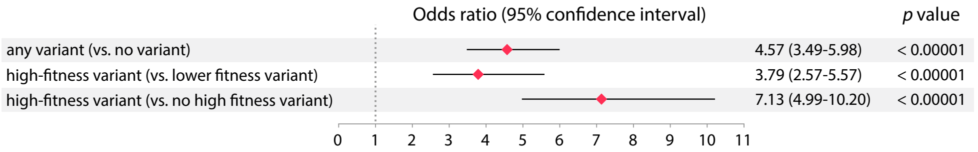


Figure 2.12 Odds ratio of AML stratified by variant fitness. Odds ratios were calculated using Fisher's Exact Test (one-sided) using combined data from Desai 2018, Abelson 2018 and Young 2019 (Appendix A.5). Only SNVs were considered when calculating the number of individuals with variants. 'High-fitness variant' refers to any of the 20 highly fit variants we identified in Figure 2.8 (Section 2.5.1). 'Lower-fitness variant' refers to any other SNV.

2.7 Age dependence of clonal haematopoiesis

Predicted prevalence of specific variants

The prevalence of a variant, within a particular range of VAFs, can be calculated by integrating the variant's probability density, given in eq. 2.1, but as a function of f = VAF, over the range of cell fractions (f_0 to f_1):

$$\int_{f_0}^{f_1} \frac{2N\tau\mu}{f(1-2f)} e^{-\frac{f}{\phi(1-2f)}} df \qquad \text{where } \phi = \frac{e^{st}-1}{2N\tau s} \tag{2.4}$$

A key prediction of our model, which assumes that variants enter the HSC population at a constant rate throughout life, is that prevalence of a specific variant is expected to increase approximately linearly at rate $2N\tau\mu s$, once the individual is above a certain age determined by the VAF limit of detection (f_{lim}) and the variant's fitness effect (s). The reason for this is that, provided the limit of detection is less than the VAF at which the exponential decline in variant densities occurs (i.e. $f_{\text{lim}} \ll \phi$), the variant's prevalence can be approximated as:

$$\int_{f_{\text{lim}}}^{f_1} \frac{2N\tau\mu}{f(1-2f)} e^{-\frac{f}{\phi(1-2f)}} df \quad \approx \quad 2N\tau\mu \log\left(\frac{\phi}{f_{\text{lim}}}\right) \quad \approx \quad 2N\tau\mu st + C \tag{2.5}$$

$$\text{where} \quad \phi = \frac{e^{st}-1}{2N\tau s} \quad \text{and} \quad C = -2N\tau\mu \log\left(2Nsf_{\text{lim}}\right)$$

We confirmed this prediction using DNMT3A R882H and R882C variants from two studies[6,75], which contained sufficient data ($>$ 500 total individuals in $\geq$ 2 age categories) to examine the age-prevalence relationship. In agreement with predictions, the age-prevalence of these variants does indeed increase linearly with age, consistent with the age-dependence of clonal haematopoiesis being driven by the expansion of clones which become more detectable at older ages (Figure 2.13).

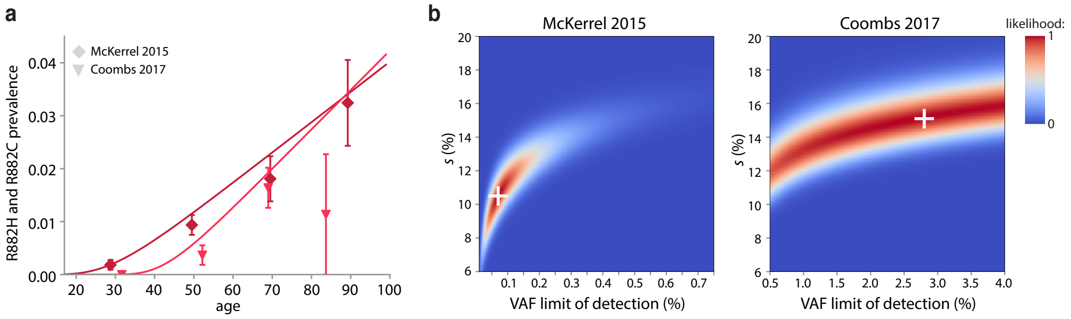
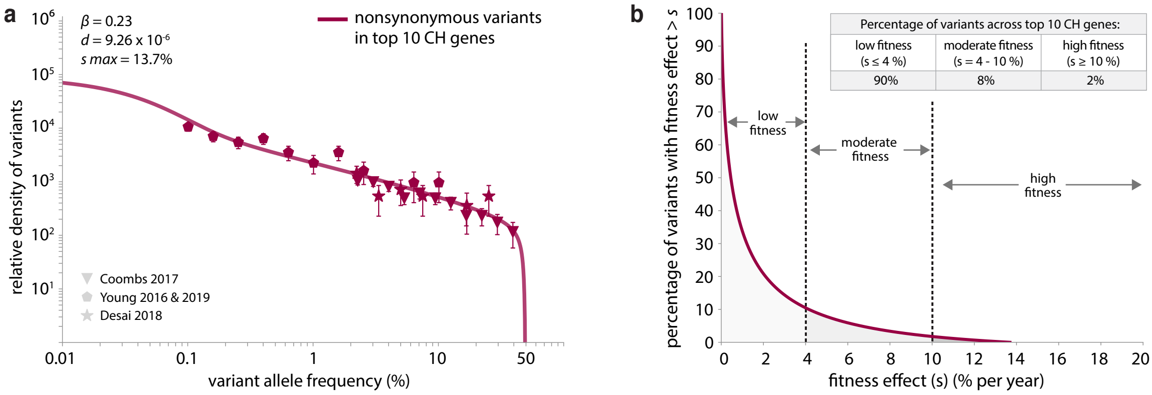


Figure 2.13 Age prevalence of R882H and R882C variants. a. R882H and R882C variants were grouped together to increase data strength. As predicted by the model, the prevalence of R882H and R882C increases approximately linearly, once above a certain age determined by the VAF limit of detection and the fitness effect of the mutation. Medium red line: McKerrell 2015, Dark red line: Coombs 2017. **b.** Maximum likelihood estimations for fitness effect (s) of R882H and R882C mutations. The most likely fitness effect (s) and VAF limit of detection were determined for the two studies that contained sufficient data to analyse their age-prevalence relationship. The white cross marks the most likely s and VAF limit of detection for the study.

Because the rate of increase in prevalence with age is $2N\tau\mu s$, examining the age-prevalence relationship of a variant also provides us with another method by which the fitness effect of the variant can be inferred. A maximum likelihood approach was used to calculate the likely fitness effect (s) and VAF limit of detection (f_{lim}), by integrating the expected density of clones (eq. 3.2) between the f_0 = VAF limit of detection and f_1 = 0.5 across a range of ages (t). The best-fit s for McKerrel 2015[6] was 10.5% per year, and for Coombs 2017[75] was 15.1% per year (Figure 2.13b), which is in good agreement with estimates inferred from the VAF distributions for R882H and R882C (Figure 2.6 and Appendix A.4).

Predicted prevalence of clonal haematopoiesis

Our framework also allows us to estimate the overall prevalence of clonal haematopoiesis as a function of age, for different sequencing thresholds. To do this, we considered the distribution of fitness effects for nonsynonymous variants across 10 of the most commonly mutated clonal haematopoiesis genes in four studies that most comprehensively targeted these genes[7,38,39,75]. Estimates for the distribution of fitness effects, s, across these 10 genes were inferred by fixing $N\tau$ to that inferred from DNMT3A R882H ($N\tau \approx$ 100,000). The distribution of ages was assumed to be Gaussian with mean 60 years and standard deviation 15 years. These are the values for the mean and standard deviation of participants from Coombs 2017[75], which contributed ~85% of the data from these four studies. We parameterised the distribution of fitness effects using a family of stretched exponential distributions (eq. 2.3) and used a maximum likelihood approach to optimise the shape (β) and scale (d) of the distribution as well as $s_{\max}$ (Figure 2.14a). This revealed a broad distribution, with 90% of variants having a low fitness and only 2% of variants having a high fitness (Figure 2.14b).



Percentage of variants across top 10 CH genes:		
low fitness (s ≤ 4 %)	moderate fitness (s = 4 - 10 %)	high fitness (s ≥ 10 %)
90%	8%	2%

Figure 2.14 Distribution of fitness effects across 10 of the most commonly mutated clonal haematopoiesis genes. a. Probability density histogram for all nonsynonymous variants in DNMT3A (all coding exons), TET2 (all coding exons), ASXL1 (exon 12), JAK2 (exon 12 and 14), TP53 (all coding exons), SF3B1 (exons 13-16), SRSF2 (exon 1), IDH2 (exon 4), KRAS (exons 2-3) and CBL (exons 8-9). **b.** Distribution of fitness effects across these 10 genes.

The prevalence of clonal haematopoiesis across these 10 genes, as a function of age, was then calculated by taking into account the distribution of fitness effects and different VAF limits of detection. Briefly, the mutation rate was normalised by the distribution of fitness effects and, for a given age, the theoretical density (eq. 3.2) was integrated over the distribution of fitness effects and over the range of VAFs capable of being sequenced (from $f_0 =$ VAF limit of detection to $f_1 = 0.499$). The predicted prevalence was then plotted as a function of age for different VAF limits of detection (Fig. 2.15). This showed that, with sensitive enough sequencing (VAFs$\gtrsim 0.01\%$), clonal haematopoiesis variants will be common, even in young adults, and almost ubiquitous in people aged over 50, consistent with recent ultra-sensitive sequencing studies[7,39].

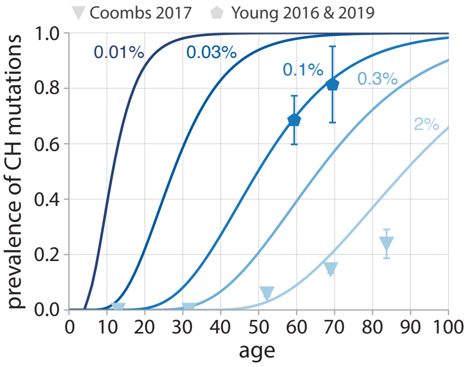


Figure 2.15 Predicted prevalence of clonal haematopoiesis as a function of age for different detection thresholds. Prevalence is predicted for individuals to have acquired at least one variant within 10 of the most commonly mutated clonal haematopoiesis genes (DNMT3A, TET2, ASXL1, JAK2, TP53, CBL, SF3B1, SRSF2, IDH2 and KRAS), taking into account the distribution of fitness effects across these genes. The actual prevalence of variants within these genes, as a function of age, is shown for Young 2016 and 2019 (pentagons, VAF limit of detection $\sim 0.1\%$) and Coombs 2017 (triangles, VAF limit of detection $\sim 2\%$. Error bars represent sampling noise.

2.8 Estimating the fitness effects of infrequently mutated sites

To determine the fitness effect of individual variants from their VAF density histograms, our framework requires that a variant be seen in $\geq$ 8-10 individuals. This means that, even with a combined study size of ~50,000 individuals, we will be unable to calculate the fitness effects of infrequently mutated variants, even if they were highly fit. A crude method to determine the fitness effects of variants is to simply determine what fitness effect, given the variant's site-specific mutation rate, would be required to explain the number of times the variant is observed. This method does not allow for deviations from the site-specific mutation rates estimated from trinucleotide context, due to its inability to separate out the effects of mutation rate and fitness (in contrast to the VAF-density histogram-based method), but it can be used to determine whether there might be highly fit, but infrequently mutated variants, that we might have missed (because they were observed in <8-10 individuals).

To crudely determine the fitness effects of all the variants that were observed more than once across a gene, e.g. DNMT3A, we first created a list of all the possible nonsynonymous variants within the gene, as well as a list of all the variants within the gene that could have been detected by each study (Appendix A.2). If a variant was included in a study's panel, the number of times the variant was expected to be observed in that study was calculated, taking into account the study's VAF limit of detection, study size, distribution of ages and the variant's site-specific mutation rate. This involved integrating the theoretical density (eq. 3.2) over the range of VAFs capable of being sequenced by the study (from f_0 = VAF limit of detection to $f_1 = 0.499$) and then integrating over the distribution of ages for that study. The expected number of observations of that variant was then summed across all the studies that included the variant in their panel. A maximum likelihood approach was then used to determine what fitness effect minimised the L2 norm between the expected and actual number of observations of each variant across all studies (Figure 2.16, Appendix A.6).

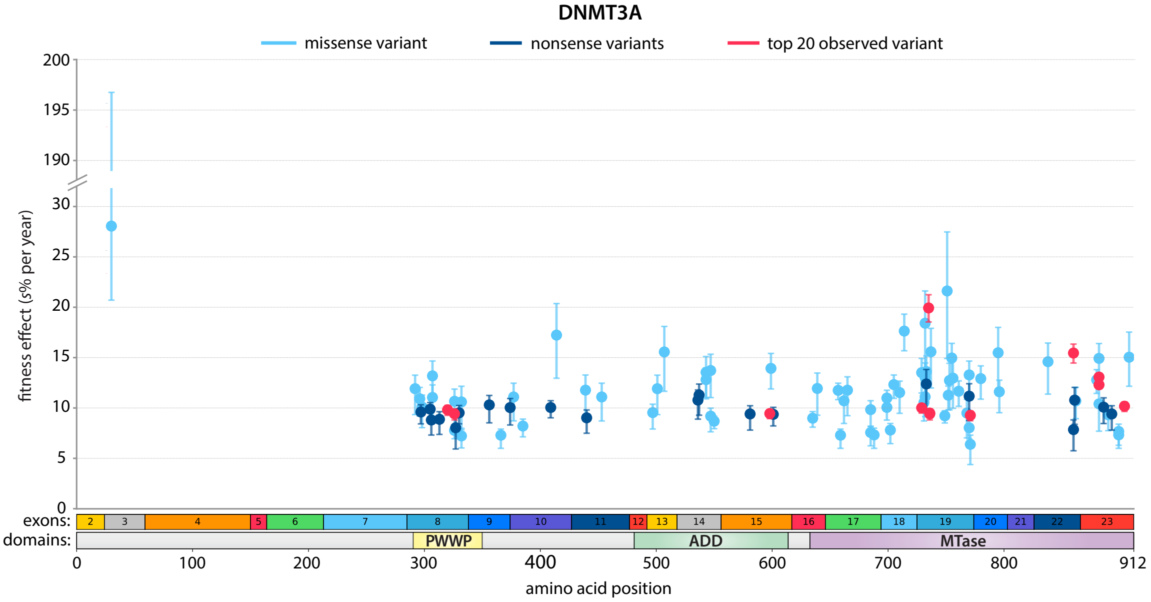


Figure 2.16 Distribution of fitness effects across DNMT3A, estimated using a counting method to infer the fitness effect required to achieve the actual number of observations of the variant. Variants that are in the list of the top 20 most commonly observed variants in clonal haematopoiesis (from Figure 2.8, Section 2.5.1) are highlighted in red. Fitness effects were calculated only for those variants observed two or more times across all nine studies. PWWP: Pro-Trp-Trp-Pro domain, ADD: ATRX-DNMT3A-DNMT3L-type zinc finger domain, MTase: Methyltransferase domain.

In addition to DNMT3A, we also used this method to determine the fitness effects of variants seen more than once, across all nine studies, in TET2, ASXL1 and TP53 (Figure 2.17, Appendix A.6). Notwithstanding the limitations caused by the inability of this method to disentangle the effects of mutation rate and fitness, this method suggests that there are a number of sites within these four mutated genes that are highly fit yet infrequently mutated.

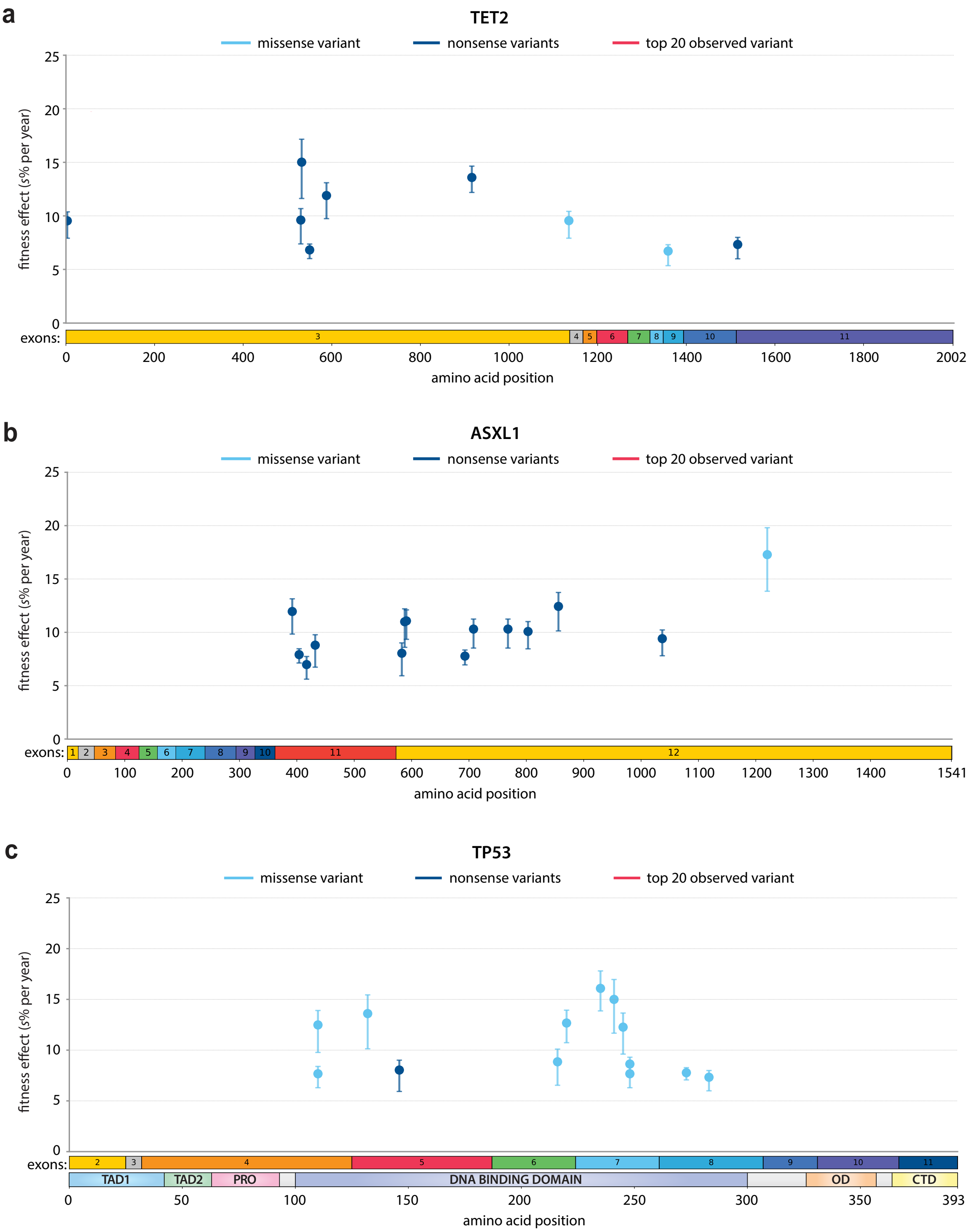


Figure 2.17 Distribution of fitness effects across TET2, AXL1 and TP53, estimated using a counting method to infer the fitness effect required to achieve the actual number of observations of the variant. No variants in TET2, ASXL1 or TP53 were in the top 20 most commonly observed variants in clonal haematopoiesis. Fitness effects were calculated only for those variants observed two or more times across all nine studies. **a.** TET2. **b.** ASXL1. **c.** TP53. TAD: Transactivation domain, PRO: Proline-rich domain, OD: Oligomerization domain, CTD: C-terminal domain.

2.9 Discussion

A simple framework explains clonal haematopoiesis

Analysing blood sequencing data from ∼50,000 individuals, using insights from evolutionary theory, shows that the VAF spectra of single nucleotide variants (SNVs) is consistent with a simple branching process model of HSC dynamics (Figure 2.18). This reveals a simple picture of how HSC population dynamics shape the genetic diversity of blood. The very wide variation in VAFs observed between people can be explained by the combined effects of chance (when a mutation arises) and fitness differences (how fast they expand). Implicit to our analysis is the assumption that clonal haematopoiesis mutations drive clonal expansions through a fixed cell-intrinsic increase in fitness. While the data across ∼ 50,000 individuals is quantitatively consistent with cell-intrinsic fitness effects playing the major role in shaping the variation in VAFs we see between individuals, it is important to bear in mind that cell-extrinsic effects such as chemotherapy[75,96–98], acute infection[99,100] and inflammation[101] likely play a role in certain contexts. Indeed, variants in certain genes (e.g. PPM1D, TP53, CHEK2 and ASXL1) have been shown to be strongly influenced by external factors[97,98,102]). It is also possible that the fitness effects of some variants may change over time, as has been recently observed for certain DNMT3A variants, whose growth rate was found to slow in older age[103].

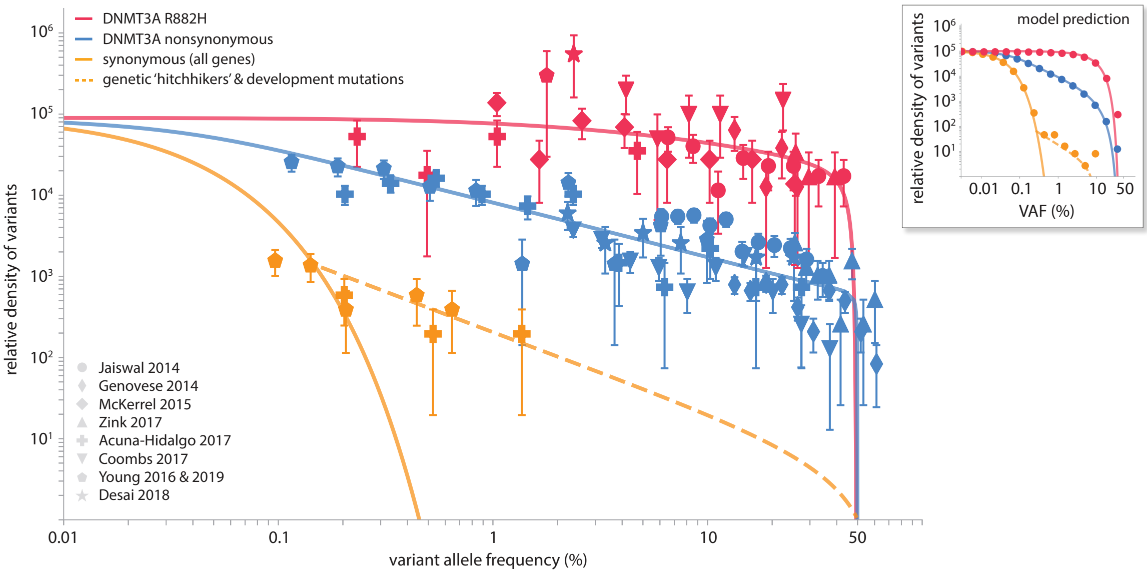


Figure 2.18 A branching model of HSC dynamics explains the observed VAF distribution for SNVs in healthy blood (Summarised from Figures 2.6, 2.10 and 2.7). The distribution of VAFs for a single variant (DNMT3A R882H, red data main plot) is consistent with the model's prediction for the distribution of VAFs for a variant with a fixed fitness effect and mutation rate (red datapoints, inset). The distribution of VAFs for a collection of variants (all nonsynonymous DNMT3A variants, blue data main plot) is consistent with the model's prediction for the distribution of VAFs for a collection of variants with a range of fitness effects (blue datapoints, inset). The distribution of synonymous variants across all genes (orange datapoints main plot) are consistent with the model's prediction for the distribution of neutral variants (orange datapoints, inset), with some of these variants hitchhiking on the back of a highly fit clone or occurring early in development (orange dashed line main plot and inset) (Hitchhiker and developmental work performed by Gladys Poon, Blundell lab[93,94]).

Whilst it may seem surprising that a simple model of HSC dynamics captures so many quantitative aspects of the clonal haematopoiesis data, work undertaken by our collaborators Alana Papula and Daniel Fisher at Stanford University has shown that more complex scenarios yield the same effective model for the multi-year development of clonal haematopoiesis, although in these scenarios N and τ have more complex meanings[93]. These more complex scenarios include models with HSCs switching between active and quiescent states, and models with progenitors occasionally reverting to HSCs. There are, however, important observations that the simple model cannot fully explain. One of these is the very broad distribution in the number of variants observed between different individuals (ranging from 0 to >10, for the same VAF detection limit[7]). One explanation for this could be variations in mutation rates between individuals or environment-specific effects. Indeed, analysis of the trinucleotide-context of mutations has revealed distinct mutational signatures for different types of cancers, some of which reflect particular mutagenic exposures (e.g. smoking, UV light or previous alkylating agent exposure), increasing age or differences in mutation repair efficiency[57–59]. Differing exposure to mutagens may result in variation in mutation rates between individuals. To investigate this further will likely require longitudinal data and would be an important area for future work.

In haematopoietic stem cells, fitness dominates drift

The relative roles of mutation, drift and selection in shaping the somatic mutational diversity observed in human tissues has been the subject of much recent debate, especially regarding the conflicting interpretations from ‘dN/dS’ measures[104–106] and clone size statistics[46,107,108]. In blood we find that the two measures are in quantitative agreement; nonsynonymous variants are under strong positive selection and most synonymous variants fluctuate via neutral drift.

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

More than 2500 variants confer moderate to high fitness

By considering the VAF spectra across ten of the most commonly mutated clonal haematopoiesis genes (Figure 2.14, Section 2.7), we can infer that mutations conferring fitness effects $s > 4\%$ per year occur at a rate of $\approx 4 \times 10^{-6}$ per year. Given the average site-specific mutation rate in HSCs is 1.6×10^{-9} per year, this implies there are $\gtrsim 2500$ variants within these genes with moderate to high fitness effects. While there is direct evidence from longitudinal data[77] and indirect evidence from age-prevalence patterns (Figure 2.13, Section 2.7) that variants at many of these sites expand at a roughly constant rate, others, notably JAK2 V617F, might exhibit more complex dynamics given the small exponential growth rates observed in longitudinal data[109]. It is likely that specific variants achieve their fitness effects in different ways. Some will simply cause a bias towards self-renewal[110,111], whereas others may cause a bias as well as an increase in the intrinsic cell division rate. Distinguishing between these scenarios is an important area for future work and will require functional studies.

DNMT3A, JAK2, TET2, ASXL1 and TP53 were the five genes with the greatest number of observed nonsynonymous variants across all the studies. We know from our analysis of the distribution of fitness effects across TET2, ASXL1 and TP53 that $\sim$2% of variants in each of these genes confer fitness effects $\geq$ 10% per year (Figure 2.10). Although this is less than in DNMT3A, in which $\sim$7% of variants are confer fitness effects $\geq$ 10% per year (Figure 2.10), it is interesting to note that we do not see any variants in TET2, ASXL1 or TP53 in our league table of the 20 most commonly observed variants in clonal haematopoiesis (Figure 2.8a). It is important to bear in mind that the variants commonly observed in clonal haematopoiesis are not necessarily the most fit, but are both sufficiently fit and sufficiently frequently mutated to be detected. Indeed, by considering all variants in TET2, ASXL1 and TP53 detected 2 or more times across all studies, we find that there are some infrequently observed variants that are potentially highly fit (Section 2.8). Thus although there are highly fit variants in TET2, ASXL1 and TP53, which are collectively not that uncommon, there were no specific individual ‘hotspot’ variants with sufficiently high enough mutation rates for them to be observed enough times for us to calculate their fitness effect.

To estimate the fitness effects of individual variants using our framework, we required the variant to have been observed in >8-10 individuals, in order to form a meaningful VAF-density histogram. The number of times a variant is observed depends on the fitness effect of the variant, the mutation rate of the variant, the size of the study and the sequencing sensitivity. Given the average site-specific mutation rate of 1.6×10^{-9} per year, a comprehensive map between variant and fitness effect for all sites that confer a fitness large enough to expand significantly over a human lifespan ($s > 4\%$) could be achieved with the current sample size ($\sim$ 50,000 individuals) by increasing sequencing sensitivity to detect variants at VAFs$>$ 0.04% (Figure 2.19). However, because sites can mutate at rates as low as $\mu \sim 10^{-10}$ / year (Table 2.2, Section 2.3.1) to quantify all variants, even rare ones, would require both a 6-fold increase in sample size, as well as sequencing sensitivities as low as 0.01% VAF (Figure 2.19). Nonetheless, even with small study sizes, there are significant advantages to being sensitive to very low VAFs[7,8,39], particularly in relation to synonymous variants, which, when grouped together, provide important information on $N\tau$ and genetic hitchhikers (Figure 2.7, Section 2.4)[94].

AML driver mutations in FLT3 and NPM1, which are found in $\sim$50% of patients with AML, were notably absent across the nine studies. One possible theory for this is that the fitness effect conferred by these mutations is so high that samples from $\sim$50,000 individuals were insufficient to capture these mutations within the brief time window between mutation acquisition and AML development. The

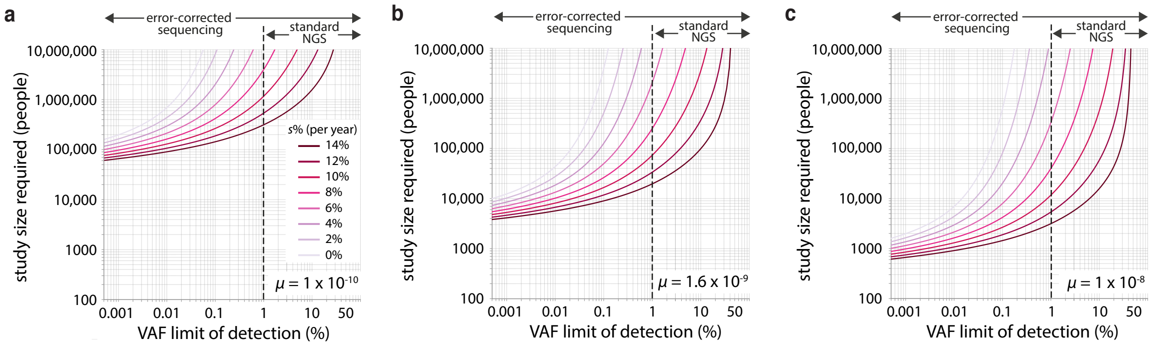


Figure 2.19 Study size required to accurately quantify different fitness effects (coloured lines) for individual variants, as a function of sequencing sensitivity (VAF limit of detection). **a.** Study size required for variants with mutation rates of 1×10^{-10} per year. **b.** Study size required for variants with mutation rates of 1.6×10^{-9} per year. **c.** Study size required for variants with mutation rates of 1×10^{-8} per year

incidence of AML with FLT3 or NPM1 mutations is ∼1 per 50,000 individuals per year and so if we assume all individuals that acquire an NPM1 or FLT3 mutation develop AML, then ∼1 in 50,000 individuals per year would be expected to acquire one of these mutations. Considering ∼50,000 individuals, with an average age of 55 (across the 9 studies), we would expect ∼ 55 individuals to have acquired an NPM1 or FLT3 mutation at some point during their life. Assuming AML develops when the VAF is close to ∼50%, the time between the mutation being detectable (i.e. >0.1% VAF) and developing AML would need to be <1 year for there to not be at least 1 person in the ∼50,000 with a detectable NPM1 or FLT3 driver mutation. A growth rate that high would involve a clone doubling time of at least every ∼ 6 weeks, which would mean these mutations confer fitness effects of at least ∼800% per year. Whilst fitness effects this high are not implausible (see Chapter 5), this theory would predict that we would see many cases of AML with FLT3 or NPM1 mutations as the sole drivers. We know from other studies that this unusual, with NPM1 and FLT3 mutations often found co-existent with other driver mutations[20]. Another potential theory for their absence in ostensibly healthy individuals is that NPM1 and FLT3 mutations on their own do not confer an unconditional fitness effect to HSCs. Indeed this theory is consistent with studies in mice and humans showing that NPM1 and FLT3 mutations are late-occurring and potentially 'cooperating' mutations that are necessary for transformation to AML[17,26].

Future directions

Clonal haematopoiesis is associated with an increased risk of cardiovascular disease[3,112] and blood cancers[3,4,38,77], and also has important consequences in the study of ctDNA[113,114], aplastic anaemia[115], response to chemotherapies[116,117] and bone marrow transplant[96,118,119]. A major challenge is to develop a predictive understanding of how variants and their VAFs affect disease risk. Recent studies show that both gene identity and VAF are predictive of progression to AML[38,77]. Our framework provides a rational basis for quantifying the fitness effects of these variants and understanding VAF variations. Using our framework we demonstrate that fitness estimates can be used to stratify AML risk. Given higher VAFs are strong predictors of AML development[38,77] and fitter variants are more likely to reach higher VAFs, it is perhaps unsurprising that high fitness variants can stratify AML risk. However, knowing the fitness effect of a variant allows prediction of which variants have the potential to reach high VAF and so should have increased predictive power, particularly when considering risk over longer timescales. Although the vast majority of individuals will never acquire the full complement of mutations required to progress to AML, combining our framework with studies that longitudinally track individuals over time will shed light on how these initiating mutations acquire further mutations that drive overt disease. More sensitive sequencing techniques, broader sampling of the genome and the study of environmental factors that alter the fitness of mutations, will also improve our quantitative understanding of native human haematopoiesis and help accelerate the development of risk predictors.

3

Fitness consequences and mutation rates of mosaic chromosomal alterations in clonal haematopoiesis

3.1 Introduction

While clonal haematopoiesis analyses have largely focused on single nucleotide variants (SNVs) and indels in recurrently mutated leukaemia-associated genes, ∼70% of clonal expansions in healthy blood are driven by mutations not covered by typical cancer-associated gene panels[94]. Mosaic chromosomal alterations (mCAs) are common in myeloid malignancies such as AML[120] and so it seems reasonable to suspect that mCAs may be driving some of the clonal expansions we observe in healthy individuals, some of whom may eventually progress to myeloid malignancy. In support of this, a number of studies have found mCAs in the blood of healthy individuals[30–33], with recent studies using UK Biobank data showing that ∼3.5% of all individuals aged 40-70 years old have a clonally expanded mCA detectable in >1% of their blood cells[34,35] . Similar to clonal haematopoiesis variants, the prevalence of mCAs increases with age[34,35,121,122] and certain mCAs are associated with an increased risk of developing myeloid malignancies (hazard ratio 17.8), at an annual incidence rate of 0.82%, and /or lymphoid malignancies (hazard ratio 28.6), at an annual incidence rate of 0.60%[36]. This risk is greater if clonal haematopoiesis variants are also present (hazard ratio ∼103) or if multiple genetic changes are present[36]. Therefore, just as the presence of clonal haematopoiesis variants in the blood has emerged as an important pre-cancerous state, so to has the presence of mCAs ('mCA-clonal haematopoiesis'). A quantitative understanding of the mutation rates and fitness effects of mCAs would aid in the development of risk stratification tools and further improve our understanding of normal haematopoiesis.

Here we adapt the evolutionary framework described in Chapter 2 to quantify the mutation rates and fitness effects of mCAs and apply this to mCA data generated by Loh et al[35] from ∼ 500,000 individuals in UK Biobank. We generate a league table revealing the fittest and potentially most pathogenic mCAs and find that chromosomal losses, as a class, tend to be the most fit. Whilst copy neutral loss of heterozygosity events (CN-LOH) are the least fit, they have the highest mutation rates. We find correlation between mCA fitness and blood cancer risk. In contrast to the strong age dependence observed in SNV prevalence in blood, we find mCA age dependence to be more variable for some mCAs, suggesting the risk of acquisition and/ or expansion of certain mCAs may be non-uniform throughout life and may be influenced by gender-specific factors.

3.2 mCA data from ∼500,000 UK Biobank participants

To estimate the fitness effects and mutations rates of mCAs, we analysed cell fraction estimates of autosomal mCAs generated by Loh et al from 482,789 UK Biobank participants aged 40-70[35]. Loh et al transformed genotyping intensities from the UK Biobank SNP array data into $\log_2$ R ratios (LRR) and B-allele frequencies (BAF) to obtain measures of total and relative allelic intensities respectively. Incorporating long-range phase information significantly increased the sensitivity for detection of BAF deviations, which meant mCAs at cell fractions as low as 0.7% could be detected.

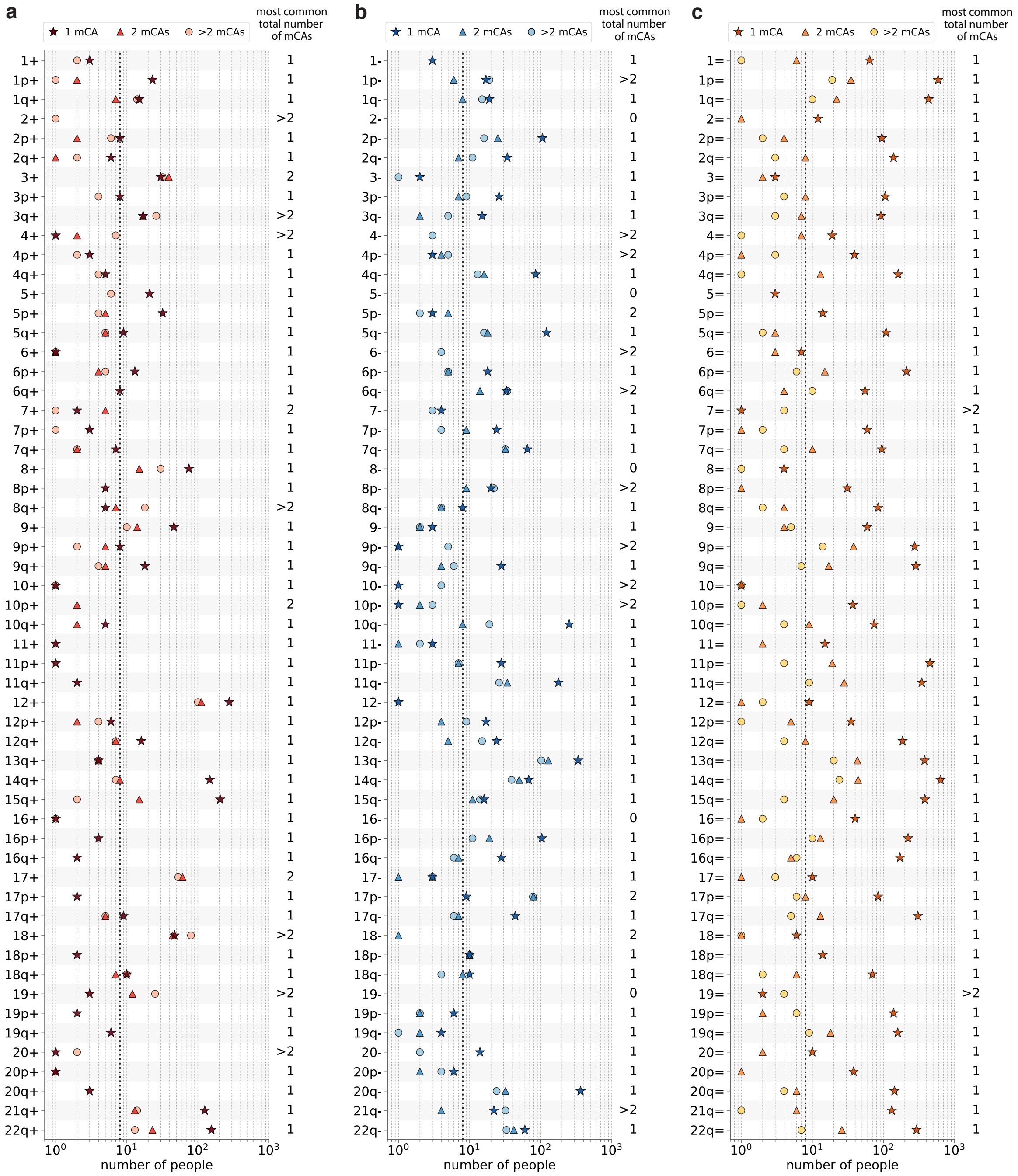


Figure 3.1 Number of observations of each mCA in Loh 2020[35], in people who had a total of 1, 2 or >2 mCAs detected. a. Gain mCAs. **b.** Loss mCAs. **c.** CN-LOH mCAs. The dashed vertical line indicates the minimum number of people (8) in whom an mCA had to be observed in order to calculate the mCA's fitness effect and mutation rate. The majority of mCAs were most commonly seen in individuals as single events ('most common total number of mCAs: 1'). mCAs that were seen more often in people that had 1 other additional mCA were 3+, 7+, 10p+, 17+, 5p-, 17p-, and 18-. mCAs that were seen more often in people that had 2 or more additional mCAs were 2+, 3q+, 4+, 8q+, 18+, 19+, 20+, 1p-, 4-, 4p-, 6-, 6q-, 8p-, 9p-, 10-, 10p-, 21q-, 7= and 19=. 6 mCAs were never seen as single events : 2+, 17+, 4-, 6- and 18-. 5 mCAs were not observed at all: 2-, 5-, 8-, 16- and 19-.

In total, mCAs were identified in 3.5% of individuals (17,111 out of 482,789): 2389 gain events, 3718 loss events and 8185 CN-LOH events (Appendix B.1). Some mCAs were observed far more often than others, with some being detected hundreds of times (e.g. 14q CN-LOH) and others not at all (e.g monosomy 5) (Figure 3.1). The majority of mCAs were most commonly seen in individuals as single events, although some mCAs were more commonly found in the context of additional mCAs (e.g. 17p-, 18+) (Figure 3.1). For individuals that had an mCA detected, the average number was 1 (range 1 - 22) (Figure 3.2).

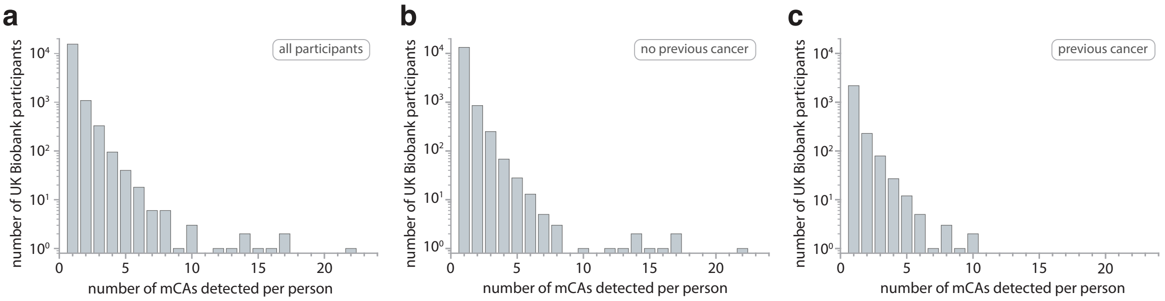


Figure 3.2 Number of mCAs per person, for individuals with an mCA detected. a. All individuals with an mCA detected (mean number mCAs = 1). **b.** Individuals with no previous cancer diagnosis that had an mCA detected (mean number mCAs = 1). **c.** Individuals with a previous cancer diagnosis that had an mCA detected (mean number mCAs = 1).

mCAs spanned a broad range of cell fractions and, as was the case with clonal haematopoiesis SNVs (Chapter 2), the density was greatest at low cell fractions (65% at cell fractions 0.7-5%) and then rapidly decreased with increasing cell fraction (Figure 3.3). There was a sharp cut-off at cell fractions $\geq$ 67% for losses and $\geq$ 54% for CN-LOH events, corresponding to BAF deviations >0.25. This was due to the analytical approach used by Loh et al[35] to identify and call mCA cell fractions, which resulted in heterozygous SNPs 'dropping out' of the data if BAF deviations were >0.25.

3.3 Evolutionary framework to infer mCA mutation rates and fitness

To disentangle how much of the variation in mCA cell fraction seen between individuals is due to differences in mutation rates versus differences in fitness effects, we adapted our evolutionary framework,[93] described in Chapter 2, to allow us to quantify the mutation rate and fitness effect of each specific mCA. The framework is still based on a simple stochastic branching model of haematopoietic stem cell (HSC) dynamics (described in Chapter 2), where mCAs (with an mCA-specific fitness effect, s) are acquired stochastically at a constant rate (μ per year), but takes in to account cell fraction (rather than VAF) measurements, the UK Biobank age distribution and the cell fraction cut-off for loss and CN-LOH events.

For a given mCA (e.g. 14q+, Figure 3.3b), cell fraction estimates were log-transformed and their density was plotted as a function of the log-transformed cell fraction, rescaled by the total number of individuals in UK Biobank. In our evolutionary framework, where the chance of acquiring an

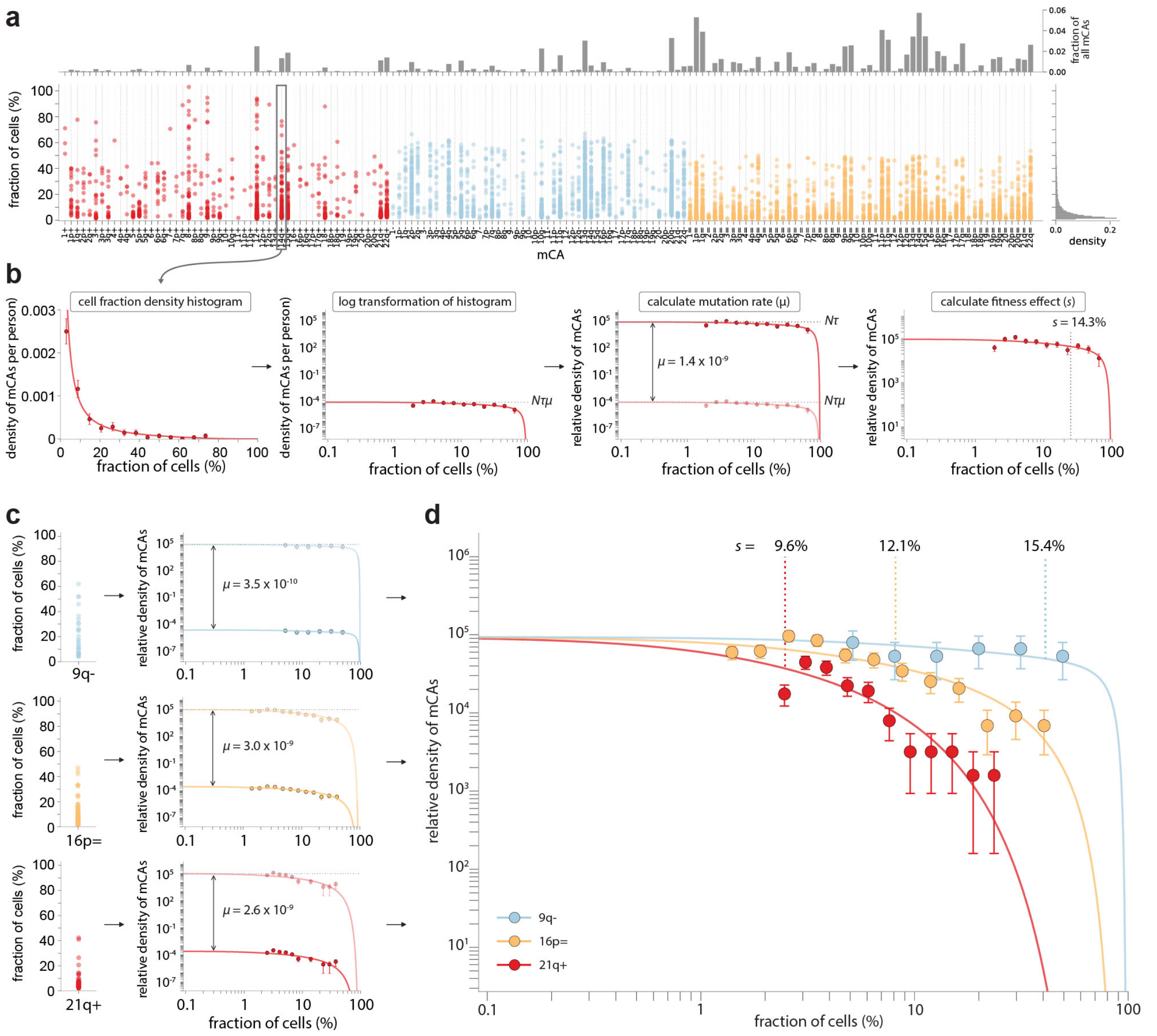


Figure 3.3 Estimating mCA fitness effects and mutation rates. **a.**. Distribution of estimated cell fractions for each mCA that was detected in ≥ 1 person in UK Biobank (red = gains, blue = losses, yellow = CN-LOH events). **b.** Plotting all cell fraction measurements for a particular mCA as log-binned histograms yields estimates for $N\tau\mu$ and s. Using an estimate for $N\tau$ of $\sim$ 100,000 allows the mCA-specific mutation rate to be calculated. Using the known distribution of ages in UK Biobank enables s to be calculated. **c.** Three example mCAs with different fitness effects and mutation rates. The mCA densities predicted by our evolutionary framework (solid lines) closely match the densities observed for specific mCAs (datapoints). The greater the fitness effect of the mCA, the faster the clone grows and so the more likely it is to be seen at higher cell fractions. Error bars represent sampling noise.

mCA is uniform throughout life, the density of a specific mCA is expected to be uniform at low cell fractions, with the amplitude determined by the product of the mCA-specific mutation rate (μ) and the haematopoietic stem cell population size. The typical maximum observed cell fraction (ϕ) for an mCA, across all individuals, is determined by how quickly the mCA-affected cells can grow (i.e. the mCA's fitness effect, s) and the longest amount of time the mCA-affected cells could have been growing for (i.e. if acquired early in life in the oldest individual). The density of mCAs is therefore expected to decline above a cell fraction determined by a combination of the mCA's fitness effect (s)

and the age distribution of individuals harbouring that mCA. How the distribution of cell fractions, predicted by our evolutionary framework, changes with age (t), the mCA-specific fitness effect (s), the mCA-specific mutation rates (μ), the population size of HSCs (N) and the time in years between successive symmetric cell differentiation divisions (τ) is given by the following expression for the probability density as a function of l = log(cell fraction):

$$\rho(l) = \frac{N\tau\mu}{(1-e^l)} e^{-\frac{e^l}{\phi(1-e^l)}} \qquad \text{where } \phi = \frac{e^{st}-1}{N\tau s} \tag{3.1}$$

Fitting the distribution of cell fractions predicted by our evolutionary framework to the observed densities for a specific mCA enables us to infer estimates for $N\tau\mu$ and s. To take in to account the varying ages in UK Biobank, predicted densities were calculated by integrating the theoretical density for a given age (eq. 3.1) across the distribution of ages in UK Biobank (23.8% aged 40-49, 33.6% aged 50-59, 42.6% aged 60-69). A maximum likelihood approach was used for parameter estimation, minimising the L2 norm between the cumulative log rescaled densities and the cumulative predicted densities, for all datapoints, in order to optimise $N\tau\mu$ and s. We now have a good estimate for $N\tau$, from work by us[93] and others[48,92], of ~100,000 and so we can estimate the mCA-specific mutation rate (μ) by dividing the maximum-likelihood inferred $N\tau\mu$ by $N\tau$ of ~100,000.

The mCA densities predicted by our evolutionary framework (solid lines, Figure 3.3c, d) closely match the densities observed for specific mCAs. Some mCAs, e.g 21q+, have a very high mutation rate, resulting in a large number of observed events, but because they only confer a modest fitness effect the vast majority are confined to low cell fraction. Others, e.g. 9q-, have a very low mutation rate, resulting in a modest number of observed events, but because they confer a substantial fitness effect, a considerable fraction are detected at high cell fraction.

3.4 Fitness effects and mutation rates of mCAs

Applying our framework to all mCAs that were observed in ≥ 8 individuals reveals a broad range of fitness effects and mutation rates (Figure 3.4a, Appendix B.2). The fittest mCAs, e.g. 3p-, 17p-, confer fitness effects in the region of $\sim$ 20% per year, which would result in a doubling of the number of affected stem cells every ~3.5 years. With this mCA fitness effect, it would take $\sim$ 50 years for an affected stem cell to clonally expand and dominate the entire stem cell pool. Therefore, even the fittest mCAs are unlikely to be detected at very high cell fraction in anyone <50 years old, unless they co-occur with other highly fit mutations. The least fit mCAs detectable in the UK Biobank participants confer fitness effects of $\sim$ 6-10 % per year, which would result in a doubling of the number of affected stem cells every ~7-12 years. This means that a stem cell acquiring one of these lower fitness effect mCA would be unlikely to expand to comprise $\gtrapprox$ 10% of the entire stem cell pool over the course of a human lifespan.

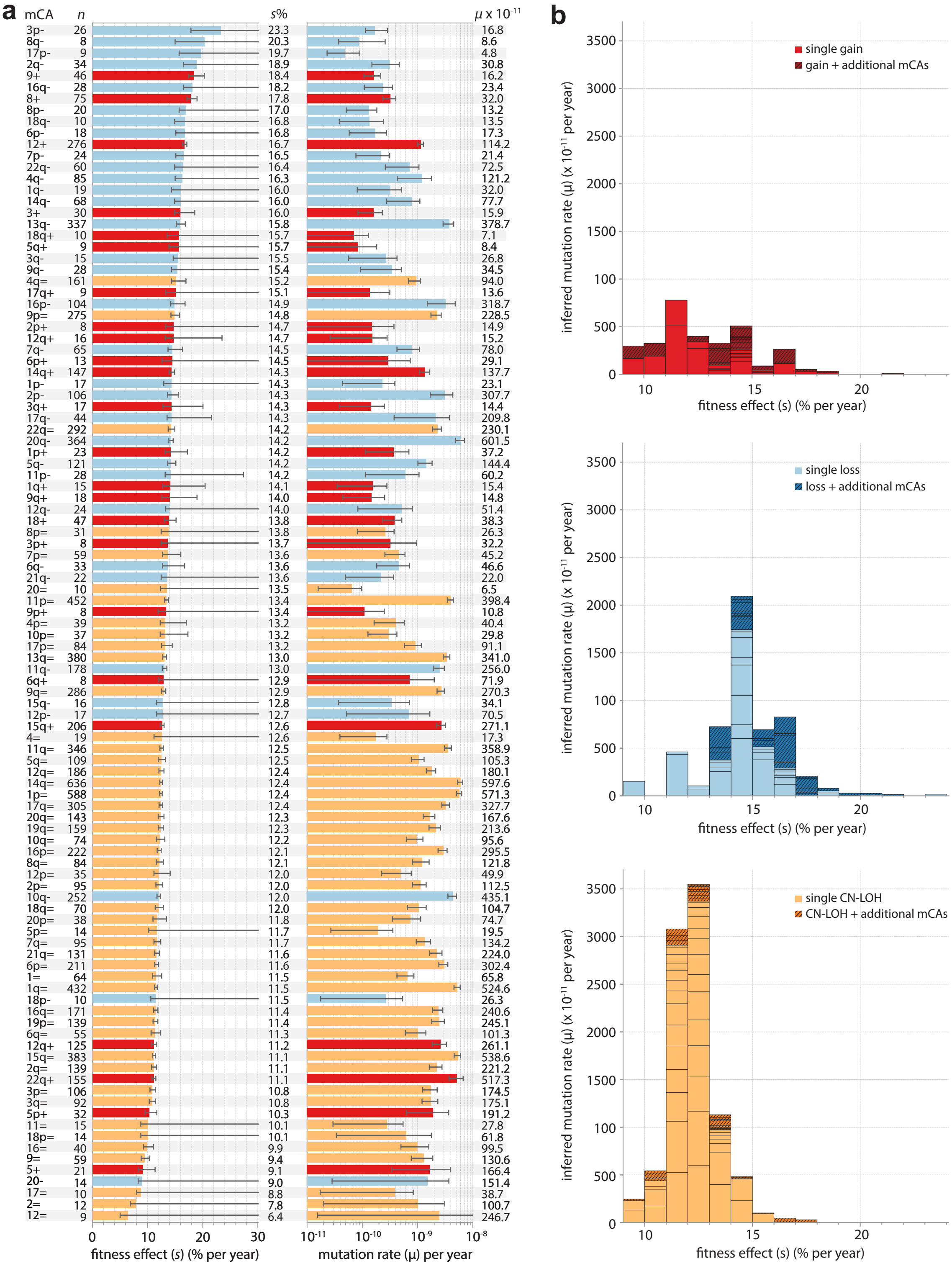


Figure 3.4 mCA fitness effects and mutation rates. a. Inferred fitness effects and mutation rates for all mCAs observed in ≥ 8 individuals. Error bars represent 95% confidence intervals. **b.** Mutation rate distribution of fitness effects for gains (red, top plot), losses (blue, middle plot) and CN-LOH events (yellow, bottom plot). Each box within a fitness interval column represents a specific mCA. Darker hatched boxes represent the fitness effects of a specific mCA that was seen in individuals that also harboured ≥ 1 other mCAs.

Examining the mutation rate distribution of fitness effects for each class of mCA reveals systematic differences between the 3 classes of mCA (losses, gains and CN-LOH events) (Figure 3.4b). Of the 3 classes, CN-LOH events occur at the highest rate (combined rate of $\sim 9 \times 10^{-8}$ per cell per year) (Figure 3.4b, bottom plot). However, CN-LOH events typically confer modest fitness effects, with most being in a narrow range between ∼11-13% per year. By contrast, the fitness effect of losses are systematically higher, with most fitness effects being between ∼14-20% per year (Figure 3.4b, middle plot). However, as a class, losses occur at a combined rate of $\sim 4 \times 10^{-8}$ per year, $\sim$ 2-fold lower than CN-LOH. Gains appear to have a broad range of fitness effects, but occur at the lowest combined mutation rate of $\sim 2 \times 10^{-8}$ per year (Figure 3.4b, top plot).

3.4.1 Sex differences in fitness effects and mutation rates

Previous studies have reported sex-biases in the prevalence of certain mCAs, e.g. 15+/15q+ is more common in men and 10q- is more common in women[34,35]. Our framework allows us to determine

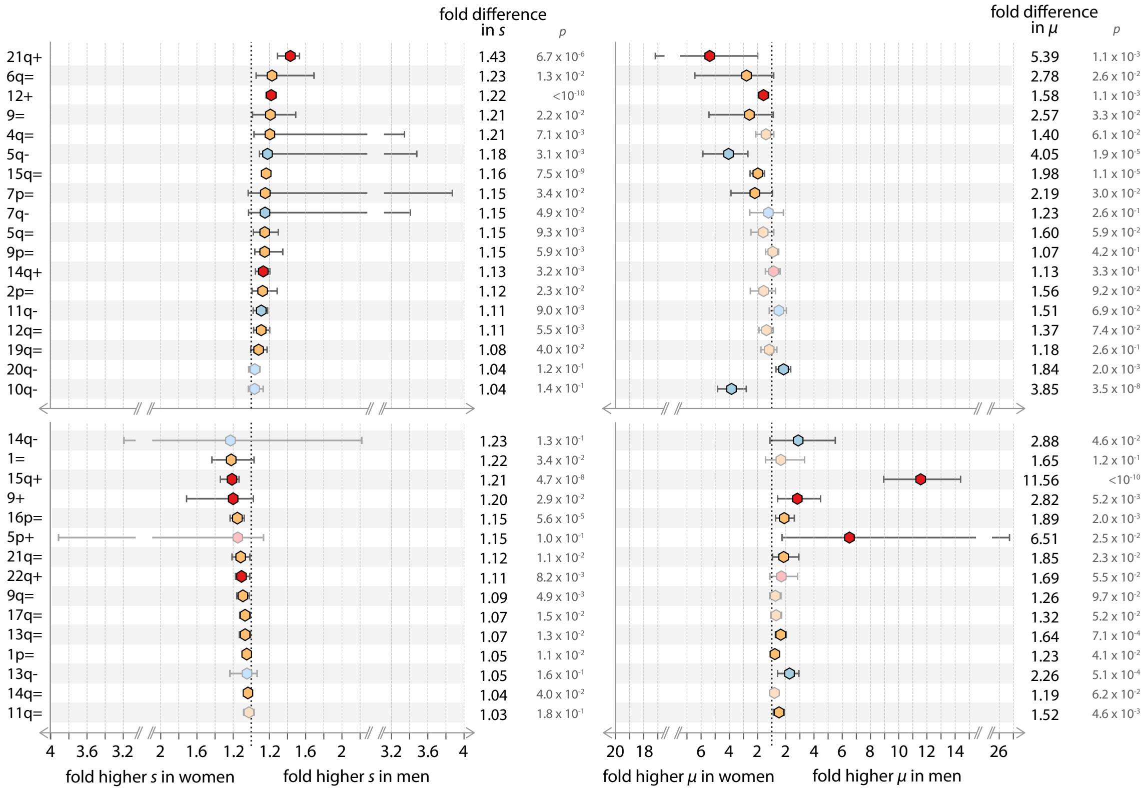


Figure 3.5 Sex differences in mCA fitness effects and mutation rates. Fold differences in fitness effects and mutation rates between men and women for mCAs that were observed as a single mCA ≥10 times in men and in women and which showed a significant difference (p-value <0.05) in either fitness effect or mutation rate. Error bars represent the 95% confidence interval from the distribution of difference between men and women. p-values were calculated from the area under the distribution of difference probability curve where the difference ≤ 0.

whether these sex-biases are driven by differences in fitness effect, differences in mutation rate, or a combination. To examine this we calculated the sex-specific fitness effect and mutation rate for mCAs that were observed $\geq$10 times in men and $\geq$ 10 times in women (Appendix B.3). The majority of mCAs (40/60) showed no significant sex-specific differences in either fitness effects or mutation rate. Of the 24 mCAs that showed significant sex differences (Figure 3.5), most sex-specific differences in fitness effect were relatively modest, with fold-differences between 1.03 and 1.43. In contrast, sex-specific differences in mutation rate were sometimes substantial, with fold-differences between 1.2 and 12.

From this analysis, we can infer that the observed higher prevalence of 10q- in women is likely due to a $\sim$4-fold higher mutation rate in women (p = 3.5 $\times 10^{-8}$), with limited evidence for any sex bias in fitness effect. Similarly, the observed higher prevalence of 15q+ in men is likely due to their $\sim$12-fold higher mutation rate (p = 0). Unlike other types of myelodysplasia (MDS), which occur more commonly in men, MDS associated with 5q- (del(5q) syndrome) is much more common in women (ratio 7:3)[123]. Our analysis suggests this may be driven by the 4-fold higher 5q- mutation rate in women (p = 1.9 $\times 10^{-5}$). Men with 5q- have a significantly higher (1.08 fold higher, p = 3.1 $\times 10^{-3}$) fitness effect, however, which may contribute to their worse del(5q) syndrome prognosis[123].

3.5 Age dependence of mCAs

The prevalence of an mCA, within a particular range of cell fractions, can be calculated by integrating the mCA's probability density, given in eq. 3.1, but as a function of f = cell fraction, over the range of cell fractions (f_0 to f_1):

$$\int_{f_0}^{f_1} \frac{N\tau\mu}{f(1-f)} e^{-\frac{f}{\phi(1-f)}} df \qquad \text{where } \phi = \frac{e^{st}-1}{N\tau s} \tag{3.2}$$

Our framework, which assumes that the fitness effects and mutation rates of mCAs remain constant throughout life, predicts how the prevalence of mCAs should increase with age. The prevalence of a specific mCA is expected to increase approximately linearly at rate $N\tau\mu s$, once the individual is above a certain age determined by the cell fraction limit of detection (f_{lim}) and the mCA-specific fitness effect (s). The reason for this is that, provided the limit of detection is less than the cell fraction at which the exponential decline in cell fraction densities occurs (i.e. $f_{\text{lim}} \ll \phi$), the mCA prevalence can be approximated as:

$$\int_{f_{\text{lim}}}^{f_1} \frac{N\tau\mu}{f(1-f)} e^{-\frac{f}{\phi(1-f)}} df \quad \approx \quad N\tau\mu \log\left(\frac{\phi}{f_{\text{lim}}}\right) \quad \approx \quad N\tau\mu st + C \tag{3.3}$$

$$\text{where} \quad \phi = \frac{e^{st}-1}{N\tau s} \quad \text{and} \quad C = -N\tau\mu \log\left(Nsf_{\text{lim}}\right)$$

We reasoned that our framework could serve as a null model to determine if there were any classes of mCA (gains, losses or CN-LOH), or specific mCAs, whose age prevalence deviated from the prevalence expected, which might highlight interesting biology.

Age dependence of gains, losses and CN-LOH events

We first calculated the expected prevalence of each class of mCA (gains, losses, CN-LOH), as a function of age. First, the expected prevalence of each individual mCA within the class (e.g. 1=, 1p= etc. for the CN-LOH class) was calculated by integrating eq. 3.2 between f_0 = mCA class-specific lower limit of detection and f_1 = mCA class-specific upper limit of detection (Table 3.1), using each mCA's sex-specific μ and s values (Appendix B.3). The overall expected prevalence for the mCA class was then calculated by summing the expected prevalence of each mCA in the mCA class. Overall, we found the observed prevalence of gain and loss events in both men and women to be in close agreement with the predicted prevalence (Figure 3.6a-c). CN-LOH events, in contrast, showed weaker age dependence than expected, particularly in women, possibly pointing to a violation of the underlying assumptions of our framework.

Table 3.1 mCA-class specific lower and upper cell fraction limits of detection. The lowest detected cell fraction for each mCA in the class, multiplied by 1.5 (to reduce the false negative rate), was calculated and the maximum of these values, across all mCAs in the class, was used as the mCA-class specific lower limit of detection.

	Gain	**Losses**	**CN-LOH**
mCA class-specific lower cell fraction limit of detection	2.5%	4.1%	1.5%
mCA class-specific upper cell fraction limit of detection	100%	67%	54%

Age dependence of individual mCAs

We next calculated the expected prevalence of each individual mCA observed $\geq$ 30 times in men and $\geq$ 30 times in women. The expected prevalence of each mCA was calculated by integrating eq. 3.2 between f_0 = mCA-specific lower limit of detection and f_1 = mCA-specific upper limit of detection, using each mCA's sex-specific μ and s values (Appendix B.3). The class-specific upper limit of detection (Table 3.1) was used as the upper cell fraction limit of detection. The lowest cell fraction detected for the mCA, multiplied by 1.5 (to reduce the false negative rate), was used as the mCA's lower limit of detection.

To quantify any deviation from the expected age dependence, the observed and expected numbers in three UK Biobank age groups (age 40-49, 50-59, 60-69) were first normalised to the observed and expected numbers in the oldest age group (age 60-69). The deviation from expected was then calculated by summing the square distance between the normalised observed and normalised expected number in each age group. By quantifying the deviation between the observed and expected prevalence across the 3 different age groups in UK Biobank we are able to identify specific mCAs with unexpected age prevalence (Figure 3.6d, Appendix B.4). While most mCAs exhibited age dependence broadly

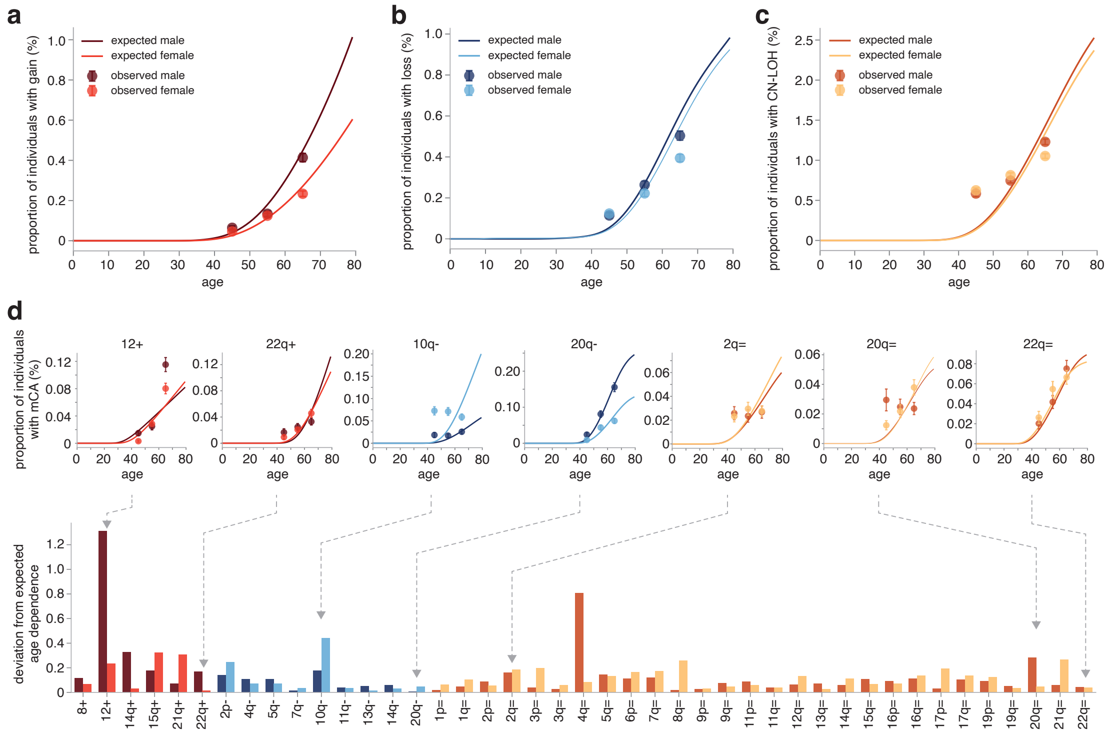


Figure 3.6 Age dependence of mCAs. a-c. Observed and expected prevalence of gains (a), losses (b) and CN-LOH (c) events for men and women. Expected prevalence (solid lines) calculated by summing the expected prevalence of each mCA in the mCA class. **d.** Deviation from expected age-dependence for each mCA observed ≥30 times in men and ≥30 times in women, with examples from each mCA class (see Appendix B.4 for age dependence plots for all mCAs).

in line with predictions (e.g. 22q+, 20q-, 22q=), there were certain mCAs that showed considerable deviation from the expected prevalence in at least one of the two sexes. Some mCAs showed greater age dependence than expected (e.g. 12+ in both men and women). Other mCAs showed no age dependence (e.g. 2q= in both men and women) and some showed declining age prevalence (e.g. 10q- in women, 20q= in men). In some mCAs, therefore, there is a lack of self-consistency between the cell fraction distributions, the age dependence and a model where the fitness effects and mutation rates are constant throughout life.

3.6 Predicted prevalence of mCAs

The observed prevalence of mCAs is determined, in part, by the sensitivity of the detection method. Because our framework predicts how the density of mCAs should be distributed as a function of cell fraction, we can predict the age prevalence of mCAs (or specific mCAs) in the blood above any defined limit of detection (eq. 3.2). With infinitely sensitive detection, the chance of an mCA being present in the blood increases steadily over the course of life, from ∼5% in teenage years to nearly 20% in later life (Figure 3.7). However, the vast majority of these are at cell fractions below the detection limit of ∼1% cell fraction in the UK Biobank dataset. The different mutation rates and fitness effects of the 3 classes of mCA drive different patterns of expected age dependence. Because losses and gains have a relatively low mutation rate, they are less common at younger ages than CN-LOH events (Figure 3.7), but because of their higher fitness their prevalence should increase more strongly with age.

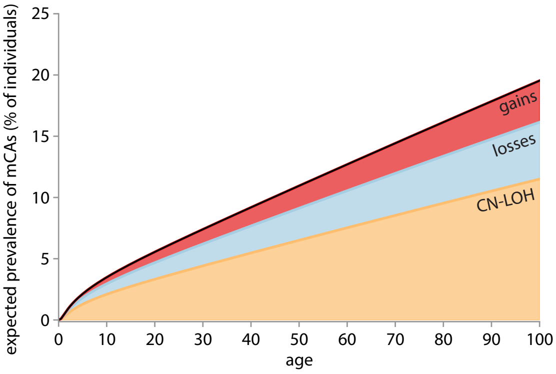


Figure 3.7 Predicted prevalence of mCAs. Predicted prevalence for each class of mCA calculated by summing the expected prevalence of each mCA (observed in ≥8 individuals) in the mCA class. Expected prevalence of each mCA was calculated by integrating eq. 3.2 between $f_0 = 1/N$ (where N ∼100,000) and $f_1 = 1.0$, using each mCA's specific μ and s.

3.7 Length dependence of mCAs

Strong clustering of loss events can be seen involving genes recurrently mutated in clonal haematopoiesis and haematological malignancies, e.g. DNMT3A, TET2, DLEU1, IGH (Figure 3.8a), suggesting the fitness effect conferred by these loss events might be attributable to the loss of one of the cell's copies of these genes. We wondered whether the fitness effects of these loss events were similar to the fitness effects we inferred for SNVs in these genes (Chapter 2) and wondered how the fitness effects and mutation rates depended on the length of the chromosomal section lost. To assess this, we separated loss events involving these genes in to broad length categories (0-3 MB, 3-10 MB, 10-30 MB and 30-100 MB) and inferred the fitness effects and mutations rates for the loss events within each length category using our evolutionary framework (Figure 3.8b, c).

Some confidence intervals were large, due to small numbers of events in some length categories (≥ 5 events required), but for the majority of loss events the fitness effect seemed to be unaffected by the length of the loss, suggesting loss of the recurrently mutated gene was the main driver of the fitness effect (Figure 3.8b). In further support of this, the fitness effects of losses involving DNMT3A, TET2 and ASXL1 were broadly consistent with the fitness estimates we had previously inferred for SNVs in these genes (Chapter 2). Interestingly, the fitness effects of loss events on chromosome 20, involving ASXL1 and/or L3MBTL1, appeared to decrease for loss lengths >30 MB, suggesting the additional loss of a gene (or region) at the telomeric end of chromosome 20 might be having a negative effect on the fitness effect. There was not a consistent pattern for how the mutation rate varied for different lengths of loss involving these genes. With increasing length of loss, the mutation rate seemed to decrease for some genes (e.g. DNMT3A, DLEU2), but seemed to increase for others, e.g. ASXL1.

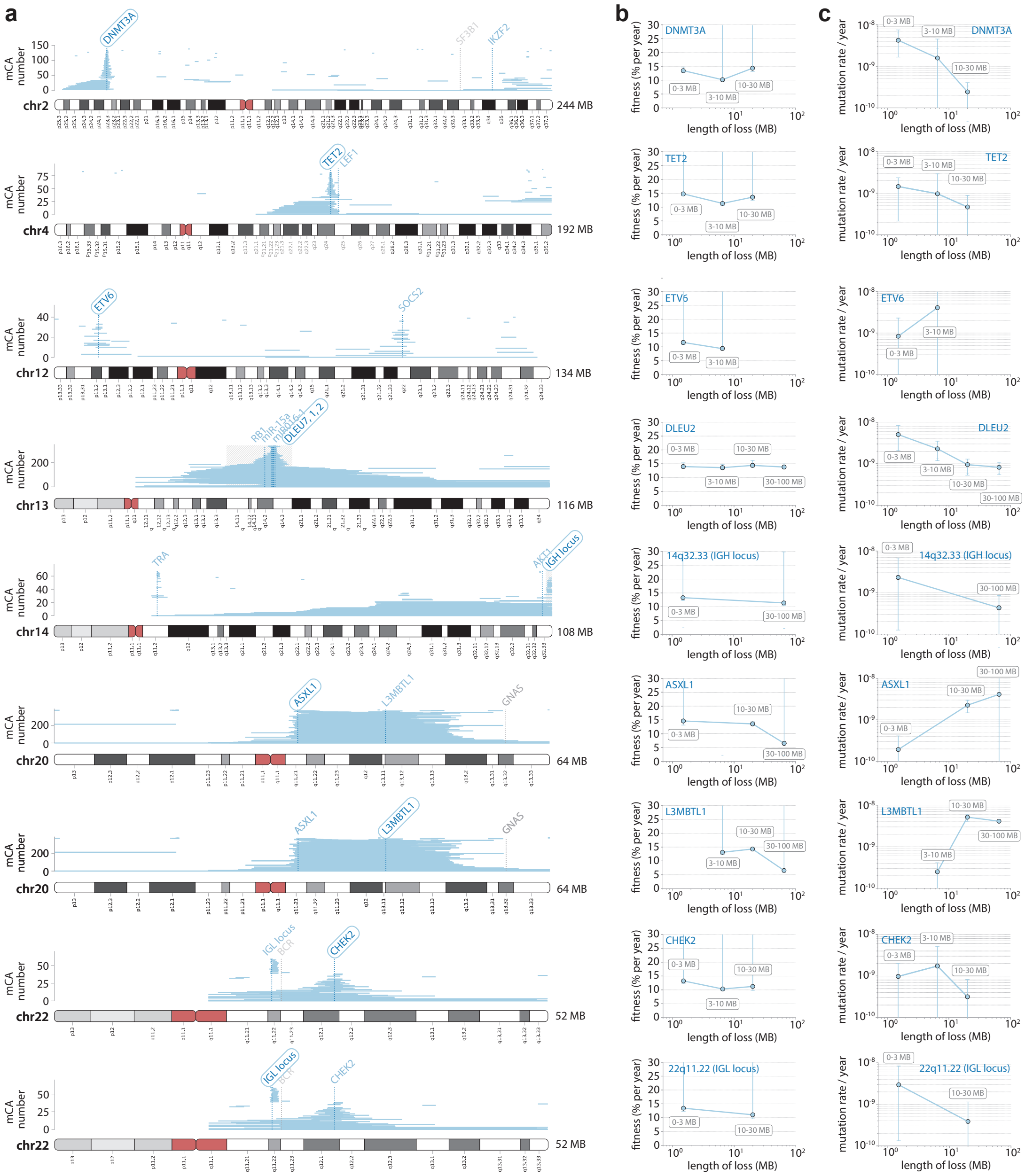


Figure 3.8 Length dependence of fitness effects and mutation rates for loss events. a. Strong clustering of loss events involving genes commonly mutated in clonal haematopoiesis and haematological malignancies was observed. **b.** Fitness effects were calculated for all losses that involved the particular gene highlighted in (a), separated into broad length categories. Error bars represent 95% confidence intervals. **c.** Mutation rates were calculated for all losses that involved the particular gene highlighted in (a), separated into broad length categories. Error bars represent 95% confidence intervals.

3.8 mCAs and haematological cancer risk

Just as clonal haematopoiesis SNVs and indels increase the risk of haematological malignancy, so do certain mCAs. Loh et al found 13 specific mCAs that were significantly associated with subsequent haematological malignancy diagnosis during 4-9 years of follow-up in UK Biobank[35]. Of these, the most significant were 12+, 13q- and 14q- which each conferred >100-fold increased risk of CLL, 9p= which conferred a 260-fold increased risk of MPNs, and 4q= and 7q= which each conferred >70-fold increased risk of MDS. Recent work by Niroula et al, also using UK Biobank data, grouped mCAs into classes based on whether they were specifically associated with myeloid malignancies (M-mCAs) (overall class hazard ratio = 28.9), lymphoid malignancies (L-mCAs) (overall class hazard ratio = 11.1) or both myeloid (overall class hazard ratio = 5.8) and lymphoid (overall class hazard ratio = 5.9) malignancies (‘ambiguous’ mCAs (A-mCAs))[36]. Interestingly, some of these were not found to be associated with haematological malignancy in Loh et al’s UK Biobank analysis[35]. This may be because Niroula et al had access to longer follow-up data (12 years) and given the median time to myeloid or lymphoid cancer diagnosis was $\sim$ 6-7 years[36], some of these cancer occurrences may have been missed in Loh’s analysis.

Table 3.2 **mCAs classified as being associated with myeloid malignancy (M-mCAs), lymphoid malignancies (L-mCAs) or both myeloid and lymphoid malignancies (A-mCAs) by niroula et al.**[36]

	Gains	Losses	CN-LOH
Myeloid associated mCAs (M-mCAs)	1q+ 8+ 9p+	5q- 20q-	9p= 22q=
Lymphoid associated mCAs (L-mCAs)	2p+ 3q+ 8q+ 12+ 17q+ 18+ 19+	1p- 1q- 6q- 7q- 8p- 10q- 11q- 13q- 14q- 17p- 22q-	7q= 12q= 13q= 16p=
Ambiguous mCAs (A-mCAs)			1p= 11q= 17p=

We wondered whether an mCA’s fitness effect was correlated with its risk of subsequent haematological malignancy. Comparing our inferred mCA fitness effects with the log odds ratio of any blood cancer reported in Loh et al, we find correlation between increasing mCA fitness and increasing blood cancer odds ratio (Pearson $R = 0.53$, $p = 9.8 \times 10^{-5}$), with all high risk mCAs having fitness effects >11% per year (Figure 3.9a). There were some mCAs (15 CN-LOH events and 8 loss events) that were inferred to be highly fit, but were not found to be associated with increased risk of any blood cancer (Figure 3.9a), within the 4-9 years of follow-up.

Correlation between increased fitness effect and odds ratio was also seen for CLL (Pearson $R = 0.55$, $p = 0.01$), with all mCAs conferring a CLL odds ratios > 100 having fitness effects >14% per year. For MPN, there was a trend towards increased fitness effect and odds ratio, although it was not significant (Pearson $R = 0.37$, $p = 0.24$) (Figure 3.9b). For MDS, higher odds ratios were not associated with higher fitness effects, although all of the mCAs significantly associated with MDS risk had fitness effects >11% per year (Figure 3.9d).

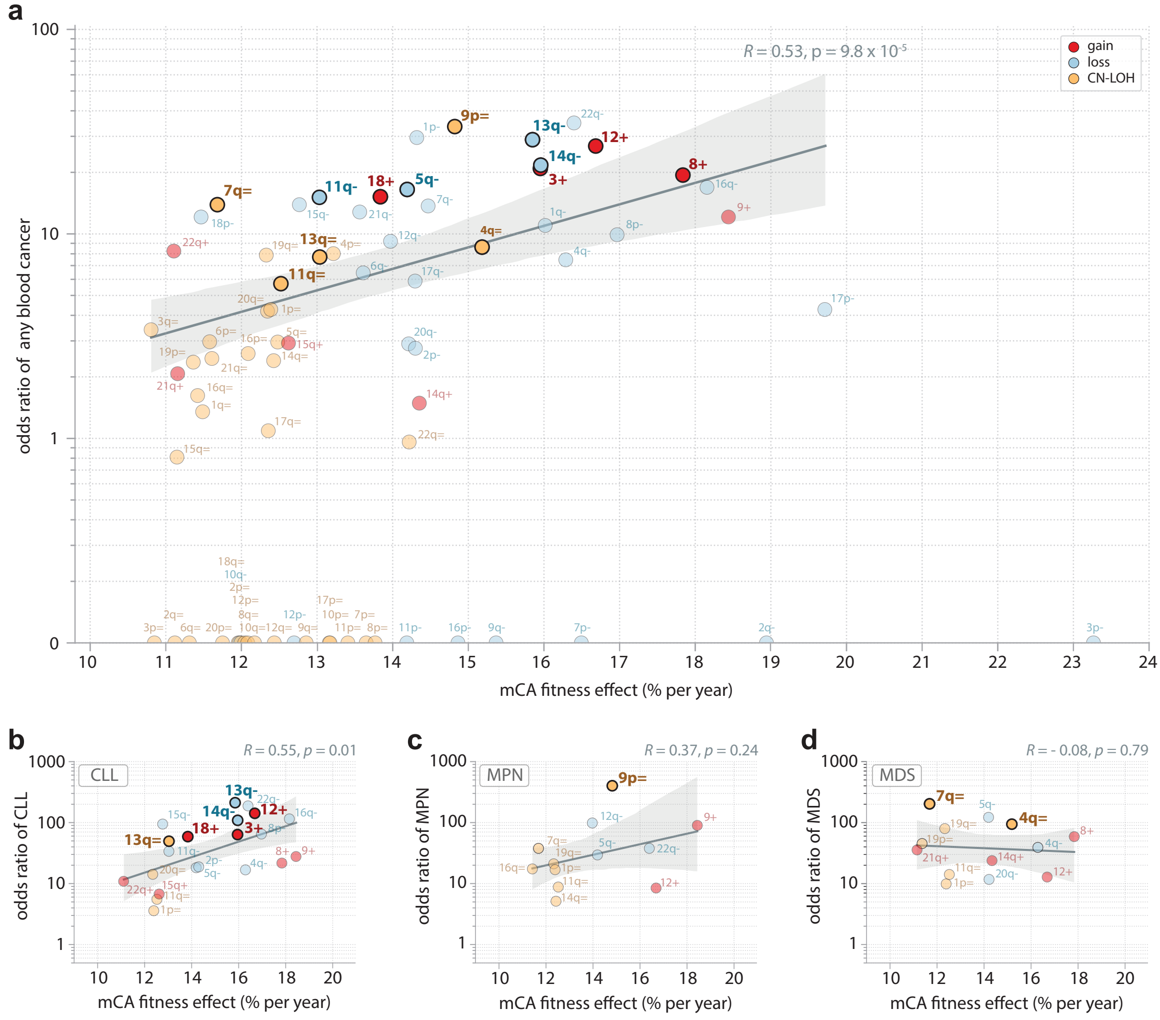


Figure 3.9 mCA fitness effects and haematological cancer risk. The odds ratios of haematological cancer for mCAs observed in ≥ 30 individuals are those reported by Loh et al[35]. The haematological malignancies were diagnosed >1 year after DNA collection (within 4-9 years follow-up) in individuals with no previous cancer. The mCAs highlighted in bold are the mCAs that Loh et al determined have a statistically significant increased risk (FDR <0.05). The mCAs faded out did not have a statistically significant increased risk. Pearson correlation coefficient and 95% confidence intervals are shown (for all mCAs with >0 odds ratio). **a.** mCA fitness effect and odds ratio of any blood cancer. **b.** mCA fitness effects and odds ratios of CLL. mCAs that had 0 odds ratio are not shown. **c.** mCA fitness effects and odds ratios of MPN. mCAs that had 0 odds ratio are not shown. **d.** mCA fitness effects and odds ratios of MDS. mCAs that had 0 odds ratio are not shown.

3.9 Discussion

A simple framework explains the behaviour of most mCAs

Analysing mCA cell fraction spectra from $\sim$ 500,000 UK biobank participants reveals that the behaviour of most mCAs, like SNVs, is consistent with a simple model of haematopoietic stem cell dynamics. In this model, mCAs are acquired stochastically at a constant rate throughout life and then expand with an mCA-specific intrinsic fitness effect. Variation in the age at which an mCA is acquired results in the observed variation in mCA cell fractions between individuals. Whilst the data are consistent with cell-intrinsic fitness effects playing the predominant role, it is likely that cell-extrinsic effects may influence the dynamics of some mCAs. Indeed, for some mCAs, we find significantly different fitness effects and/ or mutation rates between men and women, suggesting hormonal influences and/ or sex-linked genetic influences have an effect. Previous work has shown that exposure to external beam radiation therapy increases the chance of detecting an mCA (odds ratio = 1.7)[122], again suggesting a role for external influences.

Most mCAs confer high fitness effects

By considering the cell fraction spectra across individuals for each mCA, our framework enables us to quantify mCA-specific fitness effects. There are 168 different possible mCAs that could have been detected in the UK Biobank dataset (i.e. gain, loss or CN-LOH of each chromosome, p arm or q arm, excluding 13/14/15/21/22p). To infer the fitness effect of an mCA using our framework, we needed the mCA to have been observed as single events in $\geq$ 8 individuals. Despite this, we were still able to infer fitness effects for 105 of the different possible mCAs: 86% of possible CN-LOH events, 60% of possible losses and 43% of possible gains. As a class, loss events were the fittest, with most fitness effects ranging between 12-20%. CN-LOH were the least fit, with fitness effects ranging between 6-15%, whereas gains had a broad range of fitness effects ranging from 9-18%. Therefore, of all possible mCAs, we can infer that at least $\sim$60% are 'highly fit' (fitness effect $\geq$10% per year), which means if they were acquired early in life they could expand to overwhelm the bone marrow over the course of a lifetime.

It is important to bear in mind that the fitness effects we infer for some of the fitter loss events may actually be an underestimate of their true fitness. This is because the upper cell fraction limit of detection for losses was 67% (corresponding to BAF deviations >0.25), due to the analytical method used by Loh et al to call mCAs[35]. If the exponential fall-off in cell fraction densities occurred at a cell fraction greater than this ($\phi > 0.67$), then only a lower-bound on the fitness effect could be inferred (hence the large upper confidence intervals for highly fit losses in Figure 3.4). CN-LOH also had an upper limit of detection, but at a lower cell fraction of 54%. As for losses, if the exponential fall-off in cell fraction densities occurred at a cell fraction greater than this ($\phi > 0.54$), then only a lower-bound on the fitness effect could be inferred. CN-LOH events were generally less fit compared to loss events and so this was not an issue for the majority of CN-LOH events.

One of the principles underlying pre-cancerous mutation acquisition and clonal expansion is that the greater the fitness effect of a mutation, the faster the clone will expand and the more likely it is that subsequent mutations will be acquired within the same clone. Consistent with this, we found correlation between higher mCA fitness effects and increased risk of any haematological malignancy. There were some mCAs, however, that we had inferred to be highly fit, but had no reported increased risk of haematological malignancy over 12 years of follow-up, e.g. 3p- which was observed in 26 individuals and had an inferred fitness effect of 23% per year. This suggests that additional factors, other than the fitness effect of the initial driver mutation, may be important for subsequent progression to malignancy. There is also likely to be variability in the time it takes to progress to malignancy and so 12 years of follow-up may not be sufficient to observe the subsequent development of cancer in some individuals.

mCA-specific mutation rates are similar to SNV mutation rates

Unlike somatic SNV mutation rates, which can be estimated from large-scale single-cell sequencing studies[124,125], somatic mCA mutation rates have historically been harder to calculate. Our framework allows us to do this, for individual mCAs as well as classes of mCAs, using recent estimates for $N\tau$[48,92,93]. We found mCA mutation rates ranged from 5×10^{-11} to 6×10^{-9} per year, with an average mutation rate of $\sim 1.4 \times 10^{-9}$ per year, which is similar to the average SNV mutation rate of $\sim 1.6 \times 10^{-9}$ per year (Chapter 2). As a class, CN-LOH events appear to have the highest mutation rate ($\sim 9 \times 10^{-8}$ per year) relative to losses ($\sim 4 \times 10^{-8}$ per year) and gains ($\sim 2 \times 10^{-8}$ per year).

BAF deviations, for an mCA of the same cell fraction, are greatest for CN-LOH events, followed by loss and gain events. For example, for an mCA affecting 50% of cells, the BAF deviation will be 0.25 for a CN-LOH, 0.17 for a loss and 0.1 for a gain event. This means that the sensitivity for detecting small CN-LOH events is greater than for gains and losses and explains why the limit of detection is lowest for CN-LOH events (Table 3.1). Although this could theoretically result in relative under-estimation of gain and loss mutation rates, these mCAs are generally fitter than CN-LOH events, with the typical maximum observed cell fraction much greater than the lower limit of detection ($\phi \gg$ lower limit of detection) (Appendix B.2), and so our mutation rate estimates for gain and loss events should still be robust.

Infrequently observed mCAs

Some mCAs that were not observed as single events in ≥ 8 individuals, were observed in ≥ 8 individuals as multiple mCA events (e.g. 8q+, 17+) (Table 3.1), suggesting a degree of co-operativity between some mCAs. The majority, however, were simply infrequently observed, which could be due to a low fitness effect and/or a low mutation rate. In a study the size of UK Biobank (~500,000 participants), we would only expect to observe an mCA in ≥ 8 individuals if its fitness effect was >7-9% per year, assuming it had an average mutation rate ($\mu = 1.4 \times 10^{-9}$ per year) (Figure 3.10b). However, for mCAs with mutation rates as low as $\mu = 5 \times 10^{-11}$ per year (the lowest mCA mutation rate we inferred), the mCA would need to have a fitness effect >16-18% per year to be detected in $\geq$ 8 individuals (Figure 3.10a). For mutation rates this low and cell fraction detection limits of 1.5-4%, well over 10 million participants would be needed to infer the fitness effects of mCAs of all possible fitness effects.

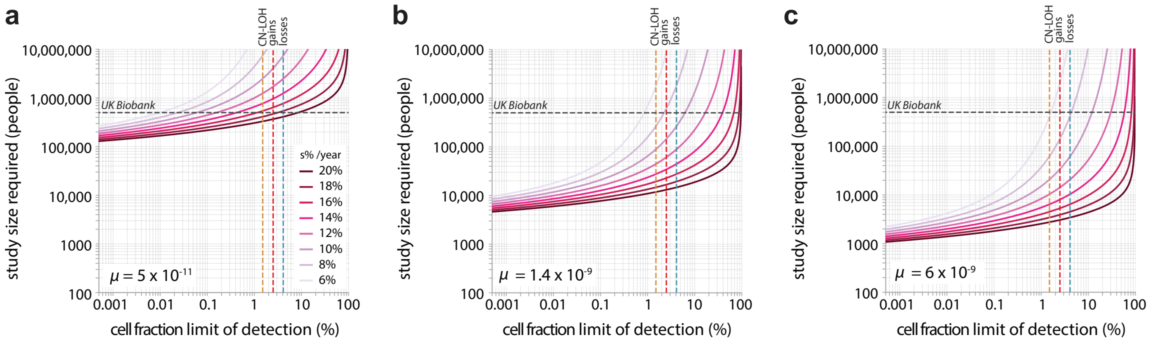


Figure 3.10 Study size required to accurately quantify different fitness effects for individual mCAs, as a function of cell fraction detection limit. Different fitness effects are represented by different coloured lines (legend in figure a). The approximate cell fraction detection limits for CN-LOH, gain and loss events are shown as orange, red and blue vertical dotted lines respectively. The UK biobank study size is indicated by the horizontal dashed black line. **a.** Study size required for mCAs with mutation rates of 5×10^{-11} per year (minimum μ inferred in Figure 3.4). **b.** Study size required for mCAs with mutation rates of 1.4×10^{-9} per year (mean μ inferred in Figure 3.4). **c.** Study size required for mCAs with mutation rates of 6×10^{-9} per year (maximum μ inferred in Figure 3.4).

There were 5 mCAs that were not observed at all in the UK Biobank dataset: monosomies of chromosomes 2, 5, 8, 16 and 19 (Figure 3.1). Of note, monosomy 5 is known to be associated with MDS and AML and is associated with poor prognosis[126,127]. Monosomy 16, although rare, has also be found to be associated with myeloid malignancies and is similarly associated with poor prognosis[128]. Whilst the absence of monosomy 5 and 16 in the UK Biobank cohort may simply reflect low mCA-specific mutation rates, their absence could suggest that these events only occur in individuals who then rapidly progress to MDS or AML (i.e. they are ‘late’ events in MDS/AML development).

Our framework highlights mCAs whose behaviour deviates from the simple model

A key feature of our framework is that it can reveal mCAs whose behaviour deviates from the simple model, thus highlighting potentially interesting biology. In contrast to the strong age dependence observed in SNV prevalence[93], we find age dependence to be more variable for several mCAs, in particular CN-LOH events. Of note, there are 8 mCAs (2q=, 3p=, 7q=, 8q=, 17p=, 20q= 21q=, 10q-) for which the prevalence plateaus or even decreases with age. This observation appears to be particularly evident in women, except for 20q= in which the observed decrease in prevalence with age only occurs in men and 2q= in which the plateau in age prevalence occurs in both men and women. The reasons behind this poor age dependence are unclear, although a number of theories are possible.

Because our mCA- specific fitness effects and mutation rates were inferred using ‘single event’ mCAs, we also focused on mCAs that were seen as single events when assessing age dependence. One theory is that the acquisition of additional mCAs with age results in the relative prevalence of single mCAs plateauing or even decreasing with increasing age. To explore this we also looked at the prevalence for individuals that had $\geq$1 mCA. However the lack of age dependence for these mCAs persisted (Appendix B.4.1), suggesting this is not the explanation.

Another theory is that these mCAs are only acquired early in life, such that above a certain age, no further increase in mCA prevalence is observed. This would suggest an age-related external factor may be important and, given this effect is largely seen in women, perhaps it could be hormonal- or pregnancy-related. Information on whether or not pregnancy has occurred is available in UK Biobank and so the potential role of this could be explored further. mCA prevalence across a large number of individuals of younger ages would also be helpful.

Another theory is that an mCA’s fitness effect could depend on the age at which it is acquired, with mCAs acquired when young being fitter than those acquired when older. Similar effects have recently been reported for DNMT3A-mutant clones whose growth is slowed in older age in the context of an increasingly competitive oligoclonal landscape[103]. In this scenario, the prevalence would be greater than expected at younger ages because more of the young mCAs, being fitter, would have expanded to above the limit of detection. These mCAs would be expected to still be detectable as the person aged, but there would be less newly acquired mCAs becoming detectable, due to their lower fitness effects. The overall result would be poorer age dependence than expected. The cell fraction density histogram for the mCA, across all individuals, would likely still appear consistent with the simple evolutionary model because the younger fitter mCAs would be interpreted as mCAs from an older individual, and vice versa. Further work is required to explore this theory further, although it does not explain why some mCAs showed a decrease in prevalence with age.

For an mCA to show a decrease in prevalence with age, either the clone needs to become undetectable, or the affected individuals cease to exist. A clone could become undetectable if it was outcompeted by a mutant-clone with a greater fitness effect. Because we focused on single-event mCAs, we know this isn’t an additional mCA, but it could be that a mutation elsewhere in the genome has occurred. This could be explored by looking at the UK Biobank whole exome sequencing data for these individuals to determine if they have additional mutations. Why these specific mCAs would be especially prone to acquiring a fitter additional clone with increasing age is not clear though. A study that analysed longitudinal blood samples showed a change in 4q CN-LOH cell fraction over time, from 58% of cells at age 82, to 59% of cells at age 88 and then 30% of cells at age 90 years[121]. A similar decrease with age was also observed in an individual with 20q-, which was detected in 51% of cells at age

75 and then only 36% of cells at age 88 years. The authors attributed these decreases with age to ‘autocorrection of the immune system’ [121]. If an mCA’s cell fraction was decreasing with age, it would need to decrease below the limit of detection to affect the overall prevalence. Whether this could be due to clonal interference or the immune system requires longitudinal data from many individuals and would be an interesting area for future work.

Whilst 10q-, 7q= and 17p= are associated with haematological malignancy (Table 3.2), the other mCAs that show poor age dependence do not appear to be associated with increased cancer risk. However, it could be that these mCAs confer a significantly increased risk of non-cancer related mortality, resulting in the affected individuals having a shorter than expected life expectancy and the prevalence of these mCAs seemingly decreasing with age. This could be explored further by exploring mortality rates in the individuals with these mCAs in UK Biobank.

Future work

Here we have shown that our evolutionary framework, based on a simple stochastic model of stem cell dynamics, allows us to infer fitness effects and mutation rates of individual mCAs and that the behaviour of the majority of mCAs, albeit with some notable exceptions (typically CN-LOH events), is consistent with this simple model. Thus far, we have focused on mCAs that were seen in individuals as single events, due to the possible confounding effect of additional mCAs on fitness effects. Multiple mCAs are not uncommon in UK Biobank, however, and some mCAs are more commonly observed in the context of additional mCAs (Figure 3.1), some of which are found at the same cell fraction suggesting they co-occurred. Exploring whether these events occur more frequently than expected by chance, and how fitness effects change with the additional of further mutations is an ongoing area we are working on.

mCAs affecting X and Y chromosomes (sex chromosomes) were not reported by Loh et al [35] and so it was only possible for us to calculate the fitness effects and mutation rates of autosomal mCAs. X chromosome mCAs have recently been associated with increased risk of lymphoid leukaemias [129] and mosaic loss of Y (mLOY), being the most common recurrent cytogenetic abnormality observed in MDS [130,131], has become important as a clonal marker in haematological malignancy. It therefore seems likely that mCAs affecting the sex chromosomes, like many autosomal mCAs, can also influence the fitness of HSCs and this would be an area for future analysis when these mCA calls become available.

In UK Biobank, mCAs >1% cell fraction were found in ∼3.5% of all individuals [35], but were found in ∼ 6% of individuals with clonal haematopoiesis variants [36] in the UK Biobank initial 50k exome release. In other studies, as many as 63% of mCAs have been found to co-occur with at least one gene mutation, which could not be simply explained by shared age-related incidence, suggesting a potential synergistic relationship [122]. Indeed, at frequently mutated DNMT3A, TET2 and JAK2 loci in UK Biobank, ∼23-60% of CN-LOH events appeared to provide a ‘second hit’ to somatic point mutations in these genes[35], with JAK2 V617F mutations being found in 60% of individuals with 9p CN-LOH events. Similarly, 7q CNLOH commonly co-localise with EZH2 mutations and 1p CNLOH events with MPL mutations[122]. Co-mutational patterns have also been observed for mCAs in *trans* with gene mutations, suggesting possible synergistic effects[122] It is therefore likely that some of our inferred mCA fitness effects may be confounded by the additional presence of gene mutations. Obtaining this data from the UK Biobank 50k exome release, and the imminent 200k whole genome release, will allow us to explore the relationship in more detail and is an important area for future work.

4

Developing an error-corrected sequencing platform for analysis of longitudinal pre-AML samples

4.1 Introduction

Distinguishing between different evolutionary scenarios is difficult using single time-point data and so, to assess whether there are differences between the evolutionary dynamics of clonal haematopoiesis in individuals who develop AML compared to 'healthy ageing' individuals, we need good longitudinal data from multiple timepoints preceding AML diagnosis. Also, if we want to study the dynamics of pre-AML mutations over time, we not only need to be able to detect a comprehensive array of AML-associated mutations, but we also need to be able to reliably detect the mutations when they are at very low frequency.

In this chapter, we describe an invaluable longitudinal blood sample resource from which we obtained blood samples, from up to 11 yearly timepoints, from 50 women who subsequently developed AML. We describe the development of a custom comprehensive targeted NGS sequencing panel, which we used to analyse these samples, which can detect an array of clonal haematopoiesis and AML associated genetic changes, including gene mutations, chromosomal rearrangements and mosaic chromosomal alterations. To detect these mutations when they are at low frequency, we used duplex error corrected sequencing and developed a custom *in silico* noise correction method, that we describe here, which allowed us to call variants down to single molecule resolution. We describe how we developed a custom chromosomal rearrangement caller for accurate translocation and inversion VAF estimation and how we harnessed the power of longitudinal samples to phase SNPs on an individual basis, enabling us to call mosaic chromosomal alterations at cell fractions as low as 0.1%.

4.1.1 Longitudinal pre-AML blood samples

The UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) is a multi-centre randomised controlled trial which recruited ∼200,000 postmenopausal women aged 50 to 74, between 2001 and 2005, to assess the impact of screening on ovarian cancer mortality[132]. All the women had a peripheral blood sample taken at trial enrolment and then ∼50,000 women were randomised to a 'multimodal' group, which involved annual blood sampling (for CA125 screening) up until 31st December 2011. All peripheral blood samples were taken in Greiner gel separation tubes and then transported overnight at ambient temperature to the central laboratory where the serum was centrifuged, aliquoted and stored. The median sample transport time (∼22.1 hours) from sample collection to centrifugation was longer than anticipated, which resulted in leukocyte DNA leaking into the serum, such that the serum contains DNA at concentrations as high as ∼ 300-750 ng/ml[133]. While this is not ideal for studies requiring 'serum samples', having high quantities of leukocyte-derived DNA in these longitudinal samples makes them an invaluable resource for our purposes.

The health outcomes of all the UKCTOCS participants are being followed until the end of 2024 and by 2018 ∼220 had developed AML (ICD-10 C92.0). Of these women, ∼50 had blood samples collected from more than 1 time-point prior to AML diagnosis (mean 5 timepoints, Figure 4.1a) and the average

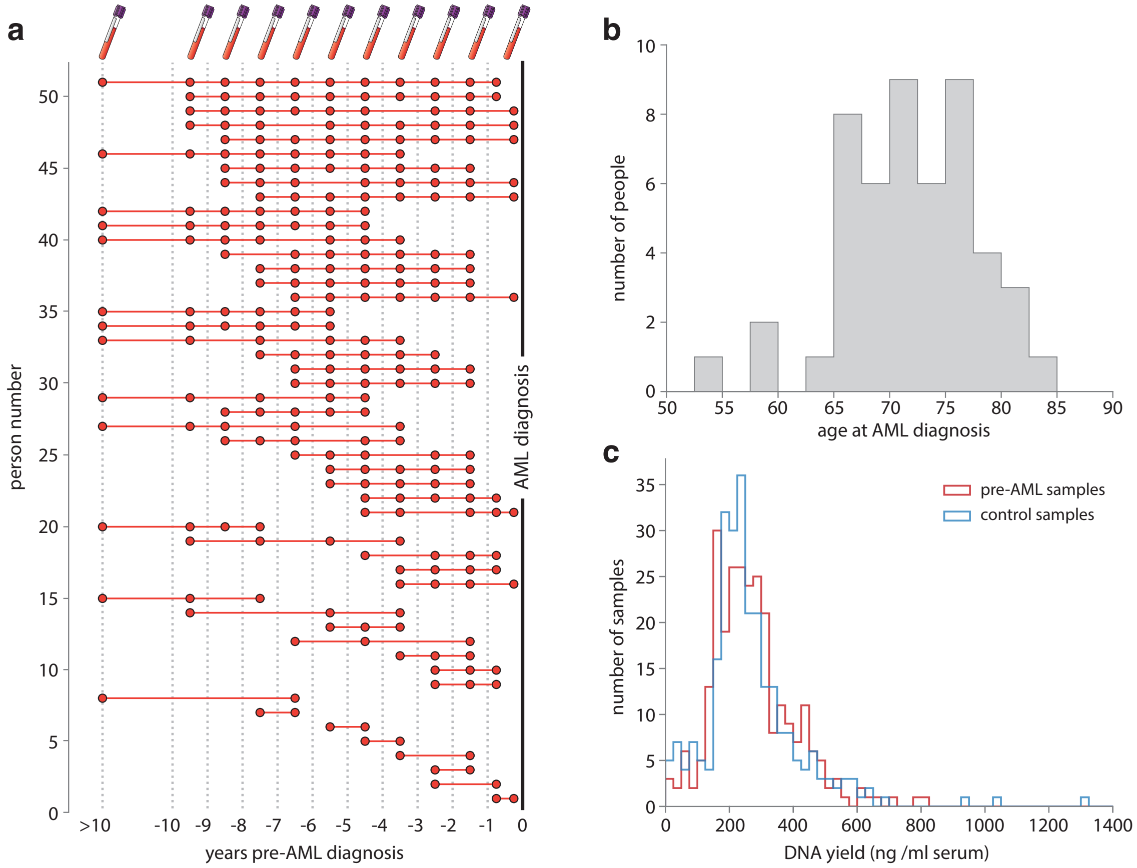


Figure 4.1 Longitudinal blood samples pre-AML diagnosis. a. Timings of blood samples in relation to AML diagnosis for all individuals in UKCTOCS who had multiple blood samples collected pre-AML diagnosis. **b.** Distribution of age at AML diagnosis for individuals in UKCTOCS: mean age 71 years, range 53-83 years, s.d. 6 years. **c.** Distribution of DNA yields from pre-AML serum samples (red histogram) and control serum samples (blue histogram). Mean DNA yield: 273 ng/ml in pre-AMLs, 269 ng/ml in controls.

age at AML diagnosis was 71 years (Figure 4.1b). We obtained 1ml serum from each of the yearly timepoints from all of the women who developed AML, as well as from age- and timepoint-matched controls who have remained blood cancer free since the trial started. DNA was extracted by LGC Genomics, using an adapted Kleargene™ method with mag beads on KingFisher™ 96, and eluted in 10 mM Tris-Cl pH 8.0. The DNA was quantified using PicoGreen™, yielding an average of 273 ng/ml DNA (range 11-814 ng/ml serum) in the pre-AML samples and 269 ng/ml DNA (range 2-1313 ng/ml serum) in the control samples (Figure 4.1c).

4.1.2 Detecting low frequency mutations using error-corrected sequencing

To trace the evolution of mutations in the longitudinal samples back in time, as early as possible, requires the ability to reliably detect mutations when they were present in just a small number of cells (i.e. $<0.1\%$ VAF). The error-rate of standard next-generation sequencing is $\sim 10^{-3}$ errors per base pair (bp) which makes it difficult to reliably detect variants $< 10^{-2}$ ($<1\%$ VAF) with standard sequencing techniques[134] (Table 4.1). However, over the past 5-10 years, much attention has been focused on strategies to improve the detection limit of sequencing, including computational / statistical strategies and various methods of 'tag-based error-correction' also known as 'error-corrected sequencing'. The principle of 'error corrected sequencing' is that each DNA molecule is 'tagged' with a unique molecular identifier (UMI) prior to amplification, so that following amplification and sequencing, all reads sharing the same UMI can be identified and grouped together. Because reads sharing the same UMI will all be copies of the same original DNA molecule, variants present in only some of the reads can be discounted as errors (e.g. due to PCR error or sequencing error), whereas those present in all reads are called as true variants. This allows variants at frequencies theoretically as low as 1.4 x 10^{-5} to $< 10^{-9}$ to be detected, depending on the error-corrected method used (Table 4.2)[134].

Table 4.1 Error rates of DNA sequencing platforms. Adapted from[134,135].

Sequencing platform	Most frequent error type	Error frequency (per bp)
Sanger sequencing	Single nucleotide substitutions	10^{-1}
PacBio RS	CG deletions	10^{-2}
Illumina MiSeq	Single nucleotide substitutions	5×10^{-3}
Illumina HiSeq 2500	Single nucleotide substitutions	10^{-3}
Illumina NovaSeq 6000	Single nucleotide substitutions	10^{-3}

Table 4.2 Accuracy of error-corrected sequencing (ECS) methods. Adapted from[134,136,137]

ECS method	Most frequent error type	Error frequency
SafeSeq	Errors due to DNA damage (e.g. 8-oxoguanine and spontaneous deamination of 5-methyl-cytosine)	1.4×10^{-5}
smMIPs	Errors due to DNA damage (e.g. 8-oxoguanine and spontaneous deamination of 5-methyl-cytosine)	2.6×10^{-5}
Duplex sequencing	Single nucleotide substitutions	$< 10^{-9}$

Error-corrected sequencing methods can be classified into one of two categories, namely 'Single-strand consensus sequencing' and 'Duplex sequencing'.

Single strand consensus sequencing

The two most common methods of single-strand consensus sequencing are SafeSeqS (or adaptations of this method)[138] and smMIPs (small molecular inversion probes)[136]. SafeSeqS was one of the first implemented methods and involves tagging the DNA molecules using PCR primers that carry UMI sequences. Following a small number of PCR cycles (first-stage PCR), further PCR amplification is carried out (second-stage PCR), but with universal primers in order to create multiple copies of each individually tagged DNA molecule. Following sequencing, all reads sharing the same UMI can be identified and grouped together[138]. The smMIPs method uses a single oligonucleotide that is made up of two arms (which bind to the target DNA region) connected via a linker region which contains a UMI. Extension and ligation between the two arms creates a tagged product which can then be enriched, amplified and sequenced and all reads sharing the same UMI can be identified and grouped together[136,139]. One of the main problems with single-strand consensus sequencing is that errors that occur during the first round of amplification will be propagated to all copies of the molecule and therefore ultimately escape error-correction. Some methods avoid this problem by ligating UMIs to the DNA molecule before PCR amplification. The main problem, however, is the mis-incorporation at sites of DNA damage (e.g. 8-oxoguanine sites and sites where there has been spontaneous deamination of 5-methyl-cytosine), which can be misinterpreted as a genuine somatic mutation.

Duplex sequencing

Duplex consensus sequencing circumvents these problems by independently tagging each strand of the DNA molecule with a UMI as well as a ‘strand-defining element’, using adapter ligation methods. This enables separate consensus sequences to be produced for each strand, but still enables the sequences to be identified as deriving from the same double-stranded DNA molecule[137,140,141]. If mis-incorporation occurs in a single strand due to DNA damage or during the first round of PCR amplification then this will be recognised as an error as it will not be present in the partner strand. With the improved accuracy of duplex sequencing, however, comes increased cost per error-corrected base, due to the need to sequence to a greater depth to achieve an adequate number of consensus sequences. The success also relies on adapter ligation efficiency, as well as the capture and amplification of both DNA strands. This means DNA: adaptor ratios and number of PCR cycles need to be tightly controlled[137].

Accepting the increased cost associated with greater sequencing depths, we decided to use a duplex sequencing approach on our longitudinal samples, as our main priority was optimising the accuracy of our low frequency variant calls. Calling variants at the single-strand consensus stage is still possible with the duplex sequencing approach and allows us the flexibility of being able to call variants at much lower VAF, due to the greater depths afforded by the single-strand consensus sequences. Although this is theoretically more ‘error prone’, we can use the single strand consensus for calling mutations of whose accuracy we can be more certain, such as variants present at very low VAF at earlier timepoints which had been called using duplex sequencing at later timepoints.

4.2 Designing a targeted sequencing panel

The heterogeneous genomic landscape of AML has now been well-characterised[20,142,143], with an average of 13 gene mutations (range 0 to 51) being found at the time of AML diagnosis, of which an average of 5 are in a set of 23 recurrently mutated genes[21]. Mutations in many of these genes are also found in clonal haematopoiesis, although some mutations, such as internal tandem duplications in FLT3 (FLT3-ITD)[20,142] and 4bp frameshift insertions in NPM1 exon 12[143], which are both suspected to occur ‘late’ in AML development, have only ever been found in individuals with AML. FLT3-ITD and NPM1 mutations are the most common mutations in AML, being found in 20-30% of individuals at diagnosis[144] and 'AML with NPM1 mutation' is now a distinct entity in the World Health Organisation classification of AML[144].

Mosaic chromosomal alterations (mCAs) are also common in AML. Approximately 67% of AML patients with intermediate or favourable risk cytogenetics have at least one detectable somatic copy number gain or loss, whilst cases with unfavourable risk cytogenetics have considerably higher numbers, with a median of 6 copy number events per person[142]. Focal deletions affecting recurrently mutated genes, such as DNMT3A, TET2, STAG2 and TP53, are common[142,145,146] and copy neutral loss of heterozygosity events (CN-LOH) are found in $\sim$ 10-35% of cases[21,147]. The presence of mCAs has historically been assessed using either cytogenetic techniques or SNP microarrays[148]. SNP microarrays typically include >1 million probes, which target SNPs evenly spaced across the genome. These probes produce two metrics for mCA detection: the B-allele frequency (BAF) and the log R ratio (LRR). The BAF is the non-reference allele fraction at a SNP locus and in a cell without an mCA will be found at either 0, 0.5 or 1.0. Gain, loss and CN-LOH events all result in deviations in the BAF. The LRR is the log ratio of signal intensity compared to a reference and is only altered when there is loss or gain of genetic material (i.e. by gain and loss events, but not CN-LOH events). Whilst SNP microarrays are classically considered the 'gold standard' for mCA detection, their mCA cell fraction detection limit, unless SNPs can be phased[35], is only $\sim$ 5-10%[30]. Over the past 10 years, a number of studies have investigated the use of next generation sequencing (NGS) for the detection of somatic mCAs[149–151], looking for deviations in heterozygous SNP VAFs and variations in read depths, in an approach analogous to SNP microarrays. These studies have shown NGS to be a sensitive and robust way of simultaneously detecting mCAs and gene mutations.

Since cytogenetic abnormalities in AML were first described in the 1970s, when t(8;21)(q22;q22) and t(15;17)(q22;q21) were discovered using chromosome banding techniques[152,153], more than 1000 balanced chromosomal rearrangements (translocations, insertions and inversions) have been identified in AML[154]. Most of the common AML-associated chromosomal rearrangements create a fusion gene which encodes a chimeric protein that is required, but not usually sufficient, for leukaemogenesis[155]. How long before AML diagnosis these chromosomal rearrangements occur is not clear. Studies in the 1990s on monozygotic twins with concordant leukemias provided evidence for a prenatal origin of some chromosomal translocations associated with childhood leukaemia. Similarly, seminal work by Greaves et al showed that chromosomal translocations, including KMT2A::AFF1and ETV6::RUNX1, could be found in archived neonatal heel-prick spots (Guthrie cards) from children who later developed leukaemia[156,157] . RUNX1::RUNX1T1 fusion transcripts have been detected in 2 of 18 healthy adult bone marrow samples and as many as 40% of cord blood samples[158]. BCR::ABL transcripts, which are more commonly associated with chronic myeloid leukaemia (CML) and acute lymphoblastic leukaemia (ALL) although can be found in AML, were reportedly found in $\sim$70% of healthy adults in a widely quoted study from 1998[159]. These findings, however, have not been replicated and their validity has been questioned[160]. Although chromosomal rearrangements are generally detected by fluorescence in-situ hybridisation (FISH), cytogenetics or real-time PCR (RT-PCR), a number of recent studies have shown that it is possible to detect them using NGS [151,161–164]. The advantage of NGS is that it enables the exact breakpoint region to be identified with single nucleotide precision and cryptic translocations, as well as rearrangements involving unknown partner genes, can be detected (e.g. KMT2A rearrangements [165]). It also does not rely on actively dividing cells, unlike cytogenetics and metaphase FISH.

To ensure that we could track the evolution of a comprehensive array of AML-associated genomic changes in the longitudinal pre-AML blood samples, we needed to design a panel that could detect gene mutations, AML-associated chromosomal rearrangements and mCAs. Similar 'comprehensive genomic' approaches have previously been published [151,164], but are typically large panels or do not include all types of genomic changes. We wanted to maximise the cost-effectiveness of our approach as much as possible and so it was important to optimise which genomic regions we targeted with our sequencing panel. Detecting mCAs involves targeting SNPs spaced across the whole genome and detecting chromosomal rearrangements involves targeting breakpoint regions which often occur in long introns. Being able to detect mCAs and chromosomal rearrangements would therefore require a large panel size (>1 MB), which would be prohibitively expensive to sequence at $\sim$ 100,000 X. However, although ultra-deep sequencing is essential for low VAF detection of SNVs and indels, it is not necessary for mCA or chromosomal rearrangement detection because the signal is spread across numerous bases. We therefore decided to develop 2 separate panels, to be used on the same input DNA sample in an integrated library preparation approach: a small panel ($\sim$ 58kb) specifically for SNVs and indels, which we would sequence at $\sim$ 50,000 - 100,000X, and a larger panel ($>$1 MB) for mCA and translocation detection, which we would sequence at a shallower depth ($\sim$ 1000 - 2000X).

4.2.1 Targeted panel for gene mutations (SNVs and indels)

To design a custom panel for targeting gene mutations, we worked with TWIST Biosciences to design a custom set of 120-nt oligonucleotide probes targeted to specific regions of interest. We chose the most commonly mutated genes from 9 different clonal haematopoiesis studies [3,4,6–9,39,75,76] as well as genes recurrently mutated in AML (e.g. NPM1, FLT3) [20,166] (Appendix Table C.1) . We then looked at the distribution of variants across these genes, in both clonal haematopoiesis and AML, and specifically targeted the exons where variants were most commonly found (Figures 4.2, 4.3, 4.4, 4.5). We also included DDX41, in which both inherited and somatic variants can be associated with an increased risk of AML [167]. Overall, the targeted panel for SNVs and indels included 32 genes and had a total size of 58,123 bp, which was covered by 589 oligonucleotide probes (Table 4.3).

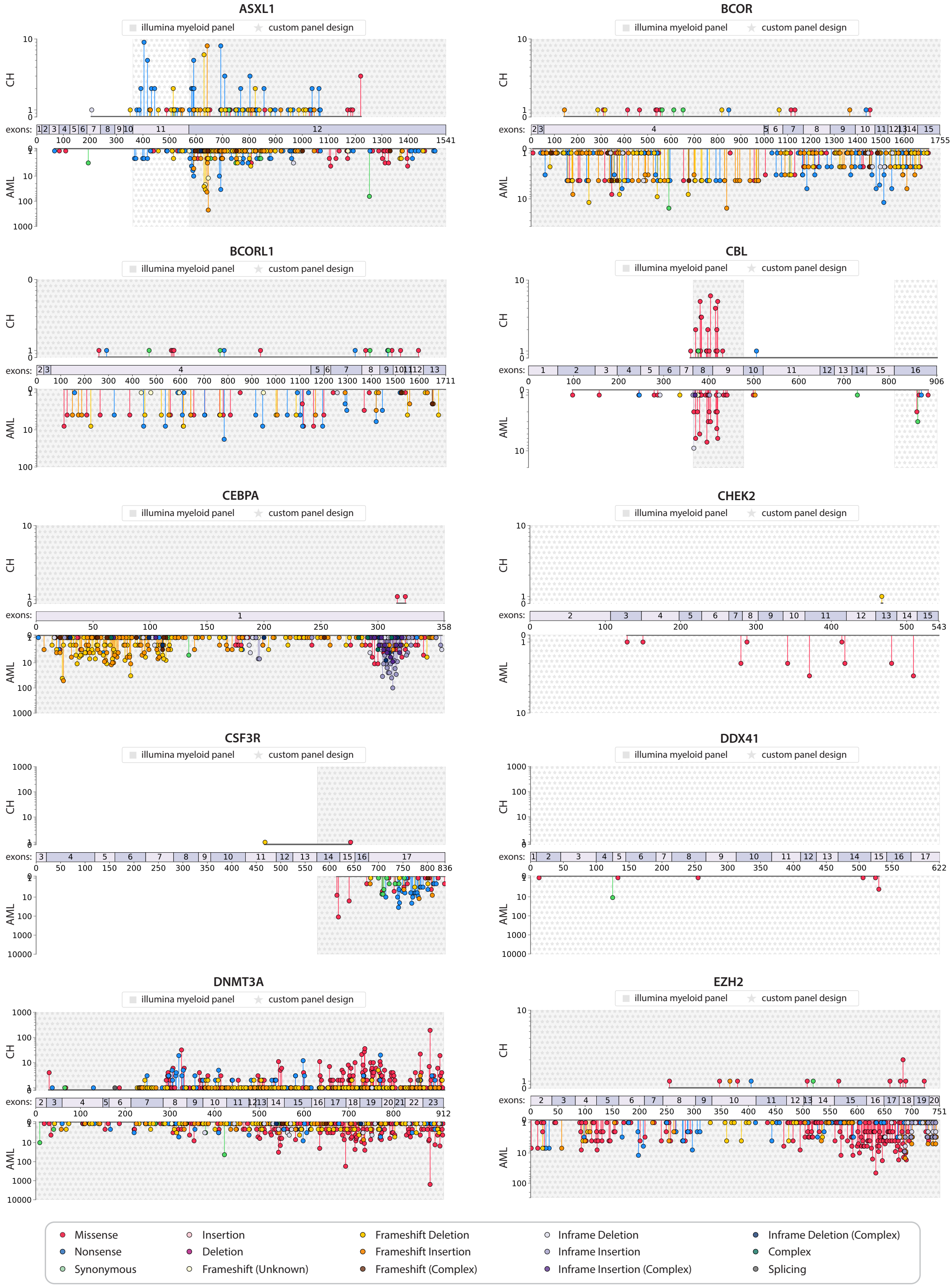


Figure 4.2 Distribution of mutations within genes in clonal haematopoiesis (CH) and AML: part 1. Regions chosen for the custom panel are highlighted with stars. Regions targeted by a typical off-the-shelf ‘myeloid panel’, e.g. Illumina TruSight® Myeloid Panel, are indicated by grey shading.

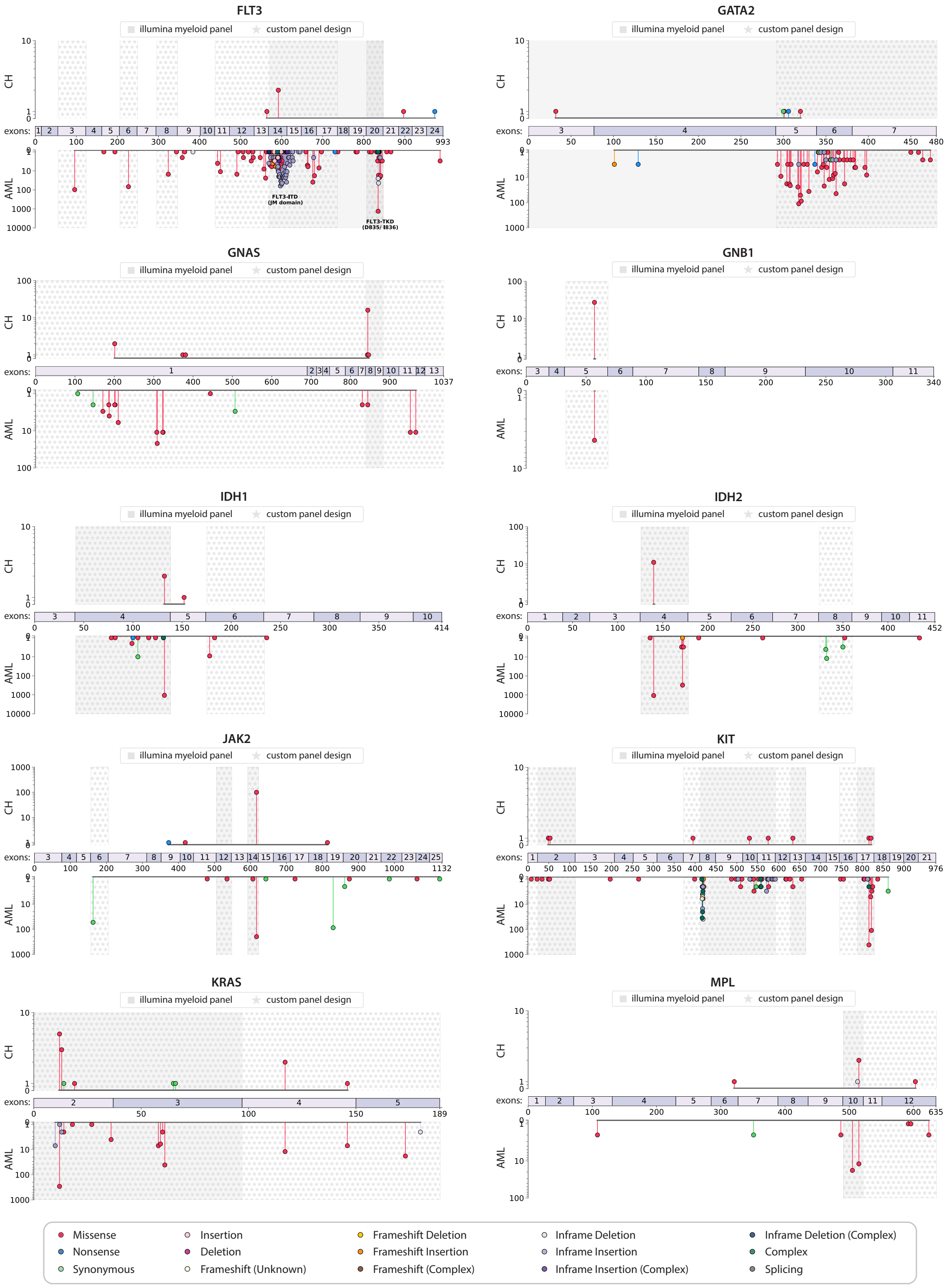


Figure 4.3 Distribution of mutations within genes in clonal haematopoiesis (CH) and AML: part 2. Regions chosen for the custom panel are highlighted with stars. Regions targeted by a typical off-the-shelf ‘myeloid panel’, e.g. Illumina TruSight® Myeloid Panel, are indicated by grey shading.

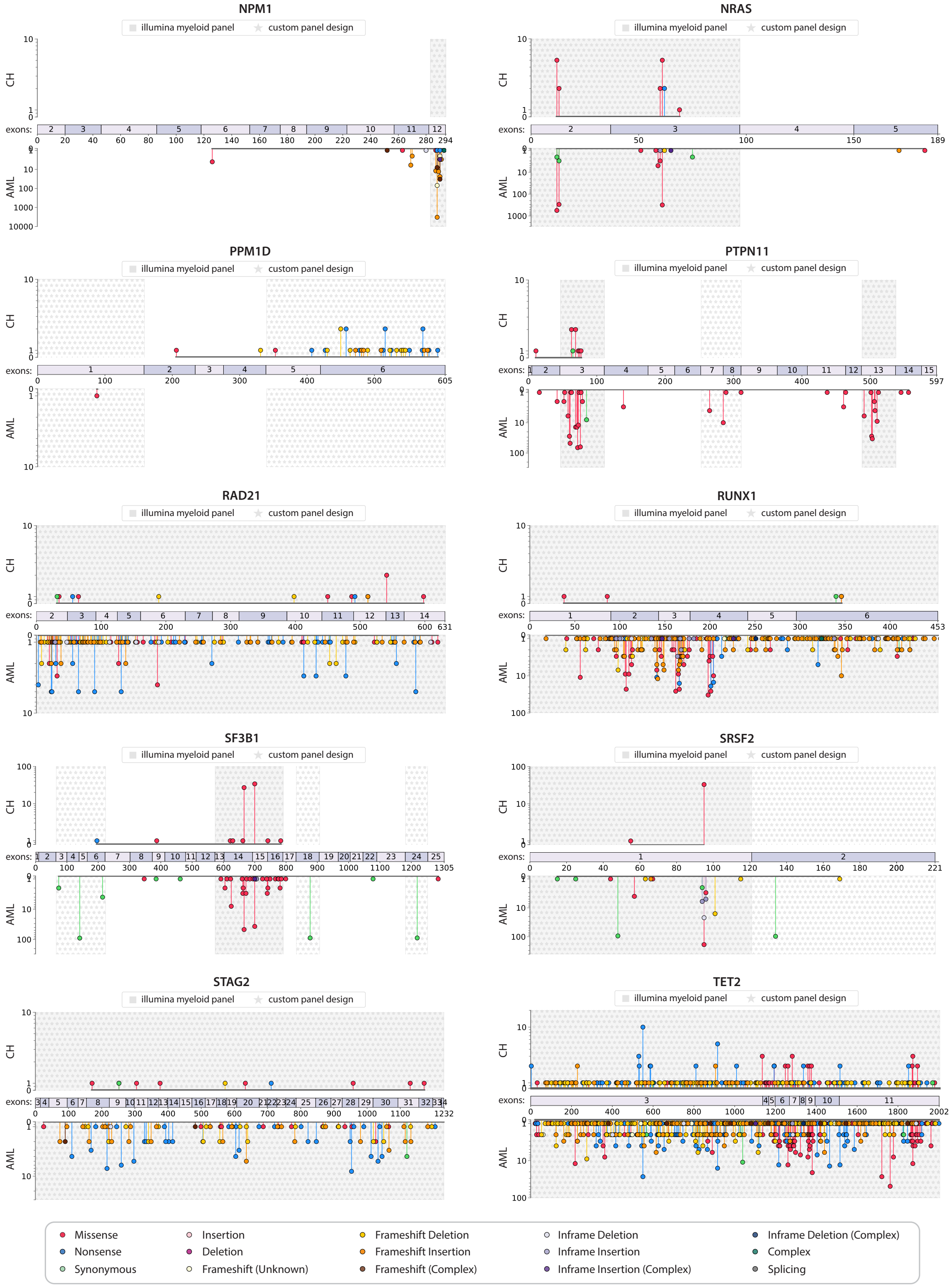


Figure 4.4 Distribution of mutations within genes in clonal haematopoiesis (CH) and AML: part 3. Regions chosen for the custom panel are highlighted with stars. Regions targeted by a typical off-the-shelf 'myeloid panel', e.g. Illumina TruSight® Myeloid Panel, are indicated by grey shading.

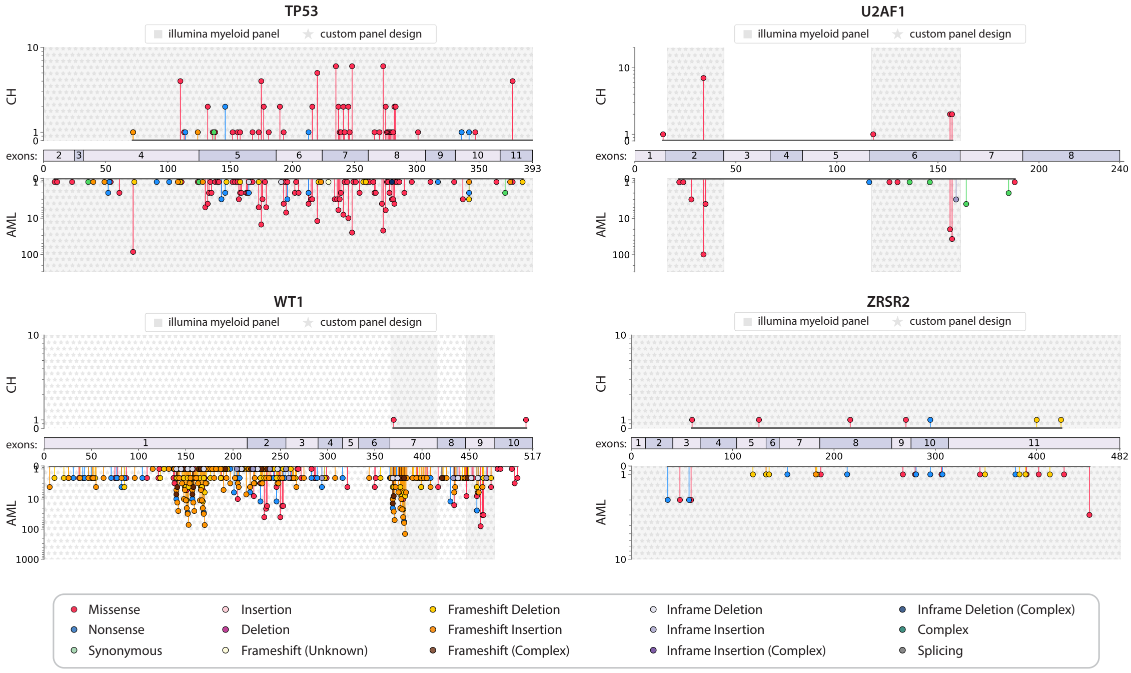


Figure 4.5 **Distribution of mutations within genes in clonal haematopoiesis (CH) and AML: part 4**. Regions chosen for the custom panel are highlighted with stars. Regions targeted by a typical off-the-shelf ‘myeloid panel’, e.g. Illumina TruSight® Myeloid Panel, are indicated by grey shading.

Table 4.3 **Gene regions targeted by custom SNV/indel panel**.

Gene	Target region (exon)	Gene	Target region (exon)	Gene	Target region (exon)
ASXL1	11, 12	GNAS	full	RAD21	full
BCOR	full	GNB1	5	RUNX1	full
BCORL1	full	IDH1	4, 6	SF3B1	3-6, 13-16, 18, 24
CBL	8, 9, 16	IDH2	4, 8	SRSF2	full
CEBPA	full	JAK2	6, 12, 14	STAG2	full
CHEK2	full	KIT	1, 2, 7-13, 16, 17	TET2	full
CSF3R	14-17	KRAS	full	TP53	full
DDX41	full	MPL	10-12	U2AF1	2, 6
DNMT3A	full	NPM1	12	WT1	1-9
EZH2	full	NRAS	2, 3	ZRSR2	full
FLT3	3, 6, 8, 11-20	PPM1D	1, 5, 6		
GATA2	5-7	PTPN11	3, 7, 8, 13		

4.2.2 Targeted panel for mCAs, KMT2A-PTD and chromosomal rearrangements

Mosaic chromosomal alterations (mCAs)

To detect mCAs, we used an approach analogous to SNP microarrays, involving the targeting of common SNPs evenly spaced across the genome (a 'SNP backbone'). This method allows us to detect deviations in heterozygous SNP B-allele frequencies (BAFs) which, when combined with read depth information, enables us to detect a gain, loss or CN-LOH event. In order to design a custom set of 120-nt oligonucleotide baits for the 'SNP backbone', common SNPs (minor allele frequency (MAF) > 0.01) were first downloaded from UCSC dbSNP (release 153, hg19). A custom Python script was then written to filter out SNPs that were likely to be uninformative. Heterozygous SNPs are essential for BAF deviations and so, to maximise their number, SNPs were filtered to retain only those whose MAF was 0.40 - 0.45 in 1000 genomes[168]. Previous work has found significant between-sample variation in normalised read depths if regions with low GC ($\leq$30%) or high GC ($\geq$60%) are targeted[150], which can make interpretation of copy number change difficult. The GC content of the region +/-60bp of each of the SNPs was therefore calculated using Pybedtools[169] and SNPs were excluded if the GC content was not between 35-55%. Mapping artefacts can result in inaccurate BAF and read depth measurements and so, to reduce the risk of this, SNPs were excluded if their surrounding region (+/-60bp) mapped to more than one location (using Bowtie2[170]) or overlapped with highly repetitive regions in Repeatmasker (hg19). A minimum of 5 SNPs are typically needed to call an mCA and so the smaller the gap between targeted SNPs, the higher the resolution for detecting shorter mCAs[150]. With better length resolution comes greater panel size, however, which will result in lower depth and therefore poorer resolution for small BAF deviations associated with low cell fraction mCAs. To strike a balance between the two, we decided to target a total of 10,326 SNPs which were spaced every $\sim$ 280 kb across the genome (Appendix table C.3), allowing us to detect mCAs as short as $\sim$ 1.5 MB.

KMT2A partial tandem duplications (KMKT2A-PTD)

Partial tandem duplications in KMT2A (KMT2A-PTD) are found in 5-10% of adult *de novo* AML and are typically associated with poor prognosis[171–173]. KMT2A-PTDs most commonly involve exons 2 or 3 and span through exon 9 to 11[171]. They can be detected by observing a relative increase in read depth, starting from exon 2 or 3, compared to an exon that is never involved in KMT3A-PTD (e.g. exon 27)[149,151]. A custom set of 120-nt oligonucleotide baits was therefore designed to target KMT2A exons 2-27 (Appendix Table C.4).

Chromosomal rearrangements

Thousands of different chromosomal rearrangements have been described in AML[154], but to limit the size of our panel we chose to focus on the detection of chromosomal rearrangements that define specific subcategories of AML in the World Health Organisation (WHO) AML classification[144]:

t(6;9) DEK::NUP214, t(8;21) RUNX1::RUNX1T1, t(9;11) KMT2A::MLLT3, t(9;22) BCR::ABL, t(15;17) PML::RARA, t(3;3) GATA2::MECOM, inv(3) GATA2::MECOM, t(16;16) CBFB::MYH11 and inv(16) CBFB::MYH11 (Table 4.4). A custom set of 120-nt oligonucleotide baits was designed to target the known breakpoint regions of each of the rearranged partner chromosomes[166,174–177]. KMT2A is renowned for having numerous possible breakpoint partners and so additional common breakpoint regions in KMT2A were also targeted[165]. Overall, coverage of the intended target regions was generally good, although the presence of highly repetitive sequences meant there were coverage gaps in some target regions (Figure 4.6, Appendix Table C.4). This should not be an issue as long as the rearranged partner's breakpoint is covered by the panel. This is because the non-targeted breakpoint region will be present as part of a 'chimeric sequence', which will be captured by the baits targeting the partner chromosome. Both breakpoint partners can therefore be identified, even if only one was targeted by the panel. The baits for targeting the breakpoint regions involved in t(3;3) and inv(3) were unfortunately designed for the wrong region of chromosome 3 (Figure 4.6) and so it will not be possible to detect t(3;3) or inv(3) with the current version of the panel.

The total size of the custom panel targeting mCAs, KMTA-PTD and chromosomal rearrangements was 1,631,472 bp ($\sim$ 1.6 MB), which was covered by 13,114 probes.

Table 4.4 Chromosomal rearrangements that define specific AML subcategories in the World Health Organisation (WHO) AML classification[144,176].

Chromosomal rearrangement	Frequency in AML	Ages	Associated genetic abnormalities	Prognosis
t(6;9) DEK::NUP214	1%	Young	FLT3-ITD mutations (78% adults).	Poor
t(8;21) RUNX1::RUNX1T1	1-5%	Young	mCAs (70%), e.g. del(9q). Mutations in KIT (20-30%), KRAS or NRAS (10-20%), ASXL1 (10%).	Good
t(9;11) KMT2A::MLLT3	9-12% (children) 2% (adults)	Any age	mCAs common (especially 8+). Overexpression of MECOM in 40%.	Intermediate
t(9;22) BCR::ABL	<1%	Adults	mCAs in most cases, e.g. del(7), 8+ or complex.	Poor
t(15;17) PML::RARA	5-8%	Any age	mCAs (40%): 8+ (10-15%). Mutations in FLT3 (30-40%).	Good
t(3;3) GATA2::MECOM inv(3) GATA2::MECOM	1-2%	Adults	mCAs common: del(7) (>50%), del(5q), complex. Mutations in NRAS (27%), SF3B1 (27%), PTPN11 (20%), GATA2 (15%), FLT3 (13%), RUNX1 (12%), KRAS (11%)	Poor
t(16;16) CBFB::MYH11 inv(16) CBFB::MYH11	5-8%	Young	mCAs (40%): 22+ (10-15%), 8+ (10-15%). Mutations in KIT (30-40%), NRAS (45%), KRAS (13%), FLT3 (14%)	Good

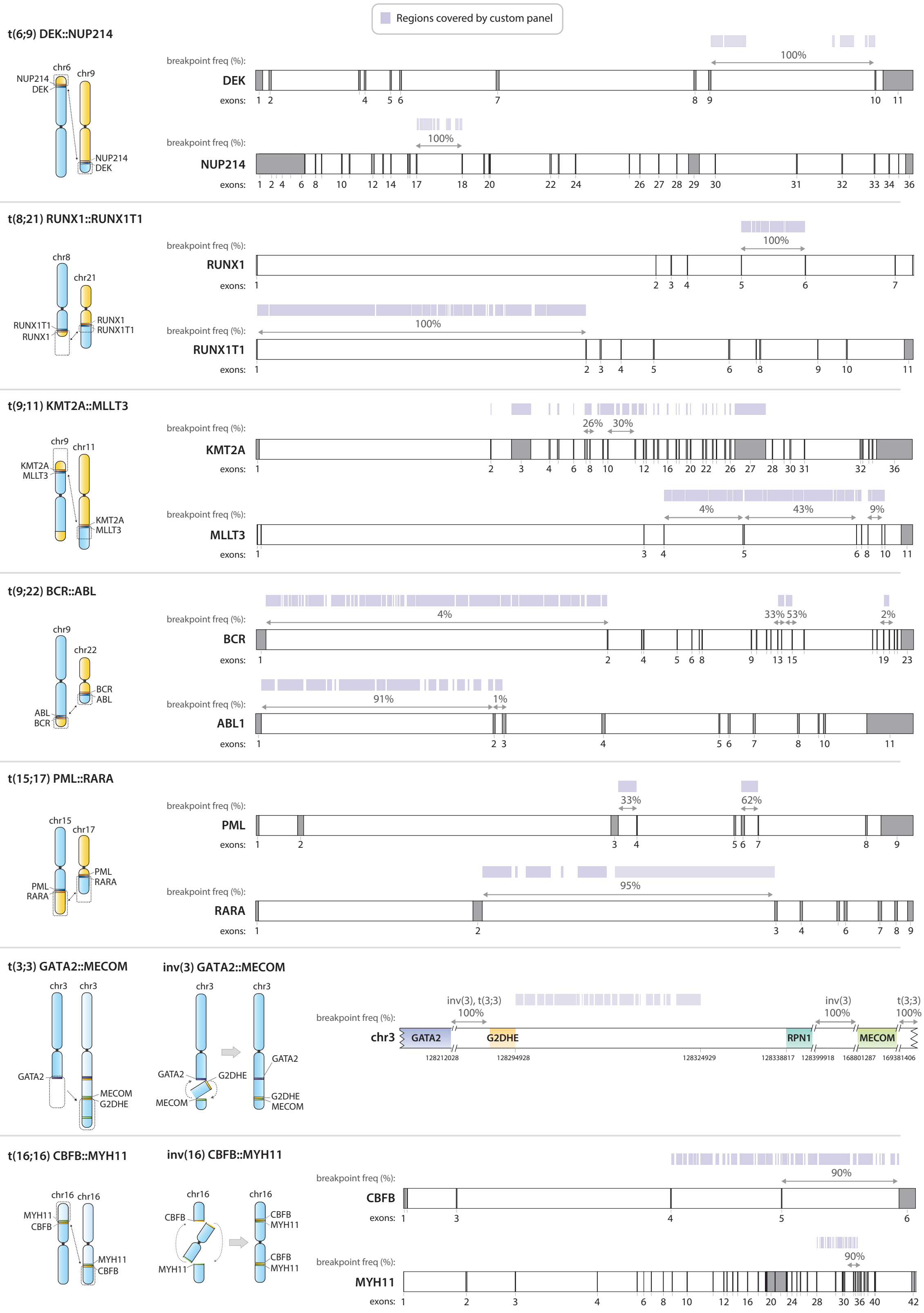


Figure 4.6 Custom panel coverage of chromosomal rearrangement breakpoint regions. The regions covered by the 120-nt oligonucleotide baits are shown in purple above the region. The frequency with which particular regions are the location of the breakpoint are written immediately above the region (indicated by arrows)[166,174–177].

4.3 Library preparation and error-corrected sequencing

Library preparation was performed as per the ‘TWIST custom panel enrichment workflow’, but with adaptations to allow for incorporation of UMIs (for error-corrected sequencing) and to allow for the use of two targeted panels on the same input DNA. DNA yield varied between UKCTOCS samples (Figure 4.1), but for each timepoint the same amount of input DNA was used for both the pre-AML sample and matched control (mean 43 ng, median 50 ng, range 4.5-80 ng).

Briefly, DNA was enzymatically fragmented (using a proprietary enzyme mix from Twist Biosciences), end-repaired and dA-tailed at 32°C for 21-24 min (see Table 4.5). IDT xGen™ CS adapters (5.5 ul of 10 μM), containing 3 bp duplex UMI sequences, were then ligated to each end of the DNA fragments at 20°C for 15 min, followed by bead purification using 0.8X DNA Purification beads (Twist Biosciences). The duplex UMI-tagged fragments were PCR amplified in a 50μl reaction volume (25μl KAPA® HiFi HotStart ReadyMix (Roche Diagnostics), 10μl IDT xGen™ unique dual index (UDI) primers (10μM), 15μl duplex UMI-tagged fragments). The following conditions were used for PCR amplification: 98°C for 45s; 11-12 cycles of 98°C for 15s (see Table 4.5), 60°C for 30s, 72°C for 30s; 72°C for 1 min. Following bead purification, using 1X DNA Purification beads (Twist Biosciences), the amplified UMI-tagged dual index-labelled fragments were eluted into 22μl ddH$_2$O and then visualised and quantified using the Agilent 2200 TapeStation.

Table 4.5 Fragmentation times and number of PCR cycles

Input DNA	Fragmentation time (min) at 32°C	1st PCR (cycles)	2nd PCR (cycles) (SNV, indel panel)	2nd PCR (cycles) (chromosomal panel)
5 ng	24	12	17	12
10 ng	24	12	17	12
20 ng	24	12	17	12
30 ng	22	11	17	12
40 ng	22	11	17	12
50 ng	22	11	17	12
60 ng	22	11	17	12
70 ng	21	11	17	12
80 ng	21	11	17	12

Following quantification, each sample was divided in two; one half for the SNV/ indel panel and the other half for the mCA/ chromosomal rearrangement panel. Samples were pooled together in groups of 8 (187.5 ng of each indexed sample), with separate pools for each of the two panels. The pooled samples were concentrated, using 1.5X Agencourt AMPure XP beads (Beckman Coulter™) and then hybridised to the custom panel probes (SNV/ indel panel probes or mCA/ chromosomal rearrangement panel probes) in a thermocycler at 95°C for 5 min followed by 70°C for 16 hours. Following Streptavidin bead capture (30 min at room temperature), the captured DNA was PCR amplified in a 50μl reaction volume (25μl KAPA® HiFi HotStart ReadyMix (Roche Diagnostics), 2.5μl TWIST amplification primers, 22.5μl captured DNA) under the following conditions: 98°C for

45s; 12 cycles (mCA/ chromosomal rearrangement panel) or 17 cycles (SNV/ indel panel) of 98°C for 15s, 60°C for 30s, 72°C for 30s; 72°C for 1 min. Following bead purification, using 1X DNA purification beads (Twist Biosciences), samples were eluted into 32μl ddH_2O and then visualised and quantified using the Agilent 2200 TapeStation. Samples were then diluted to 10 nM and submitted for sequencing.

Libraries were sequenced on the Illumina NovaSeq 6000 S4 (CRUK Cambridge Institute Genomics Core Facility) using the XP workflow, with 150 bp paired ends reads, 8 bp reads in index 1 (i7) and 8 bp reads in index 2 (i5). 40 samples were sequenced per lane with 10% PhiX control DNA spiked into each lane. SNV/indel and mCA/ chromosomal rearrangement panels were sequenced in different lanes and, with only a few exceptions, timepoints from the same individual were sequenced on different lanes. Each sequencing run contained $\sim$ 3 billion reads per lane ($\sim$ 68 million reads per sample). For the SNV/ indel panel ($\sim$ 58 kb, 589 probes), this equated to a raw depth (pre-consensus calling) of $\sim$ 50,000X. For the mCA/ chromosomal rearrangement panel ($\sim$ 1.6 MB, 13114 probes), this equated to a raw depth of $\sim$ 1000X.

4.3.1 Computational workflow for processing of sequencing data

A custom computational workflow was written for the processing of the sequencing data. This workflow used a number of software packages, including Picard[178], Fulcrum Genomics fgio package[179], Burrows-Wheeler Aligner (BWA)[180], the Genome Analysis Toolkit (GATK)[181], SAMtools[182], VarDictJava[183], Pindel[184] and ANNOVAR[185], as well as several custom written Python scripts. The workflow consisted of four main steps: i) UMI extraction and initial alignment; ii) Single strand consensus sequence (SSCS) calling; iii) Duplex consensus sequence (DCS) calling; and iv) Putative variant detection.

UMI extraction and initial alignment

Sequenced reads were demultiplexed using their sample-specific dual indexes and the demultiplexed fastq files were converted to unmapped BAM files using Picard[178] *FastqToSam*. The inline UMIs were extracted from each read and stored in the 'RX' tag of the unmapped BAM file using fgbio *ExtractUmisFromBam*. Illumina adapter sequences were marked using Picard *MarkIlluminaAdapters* and the adapter-marked BAM was converted to a fastq using Picard *SamToFastq*, which also clipped the adapter sequences. The fastq was aligned to the The Broad Institute's b37/hg19 reference genome[181] using BWA mem[180] to create a mapped BAM. BAM tags are lost during BAM to fastq conversion and so Picard *MergeBamAlignment* was used to transfer the BAM tags from the unmapped adapter-marked BAM file to the mapped BAM file to ensure that the UMI information for each read was retained.

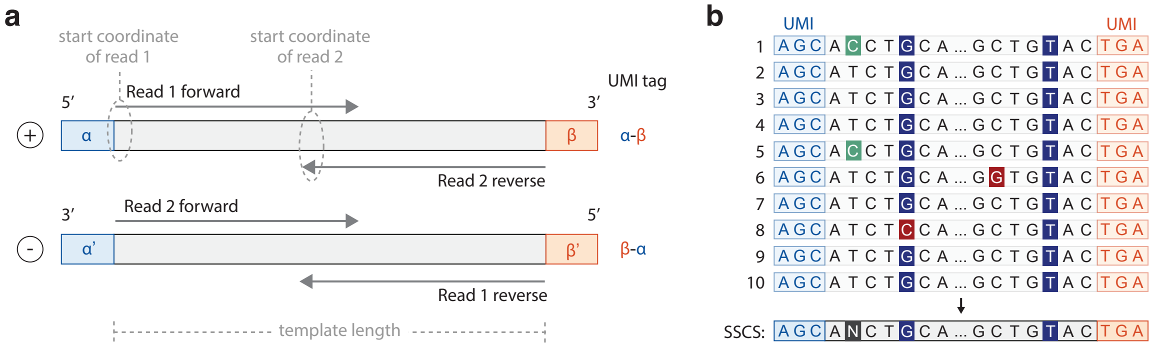

a
start coordinate of read 1
start coordinate of read 2
UMI tag
5′
Read 1 forward
3′
+
α
β
α-β
Read 2 reverse
3′
Read 2 forward
5′
-
α'
β'
β-α
Read 1 reverse
template length
b
UMI
UMI
1 AGC A C C T G C A ... G C T G T A C TGA
2 AGC A T C T G C A ... G C T G T A C TGA
3 AGC A T C T G C A ... G C T G T A C TGA
4 AGC A T C T G C A ... G C T G T A C TGA
5 AGC A C C T G C A ... G C T G T A C TGA
6 AGC A T C T G C A ... G G T G T A C TGA
7 AGC A T C T G C A ... G C T G T A C TGA
8 AGC A T C T C C A ... G C T G T A C TGA
9 AGC A T C T G C A ... G C T G T A C TGA
10 AGC A T C T G C A ... G C T G T A C TGA
SSCS: AGC A N C T G C A ... G C T G T A C TGA

Figure 4.7 UMI tags and consensus calling. **a.** Each duplex DNA fragment has a 3-bp UMI ligated at each end which is stored in the reads as a UMI tag e.g. 'α-β'. For SSCS calling, UMI families are formed by grouping reads that have the same UMI tag, start coordinate and template length and which are all e.g. 'read 1 forward'. For DCS calling, the corresponding duplex pair is identified by transposition of the UMI tag (i.e. 'α-β' and 'β-α' are a duplex pair) for reads that have the same start coordinate and template length and which are both e.g. 'forward'. **b.** SSCS calling from a UMI family containing 10 reads. Green bases represent PCR errors, red bases represent sequencing errors and blue bases represent true variants. If $\geq$90% of the bases are the same at a position, then a consensus base is called, otherwise an N is called.

Single strand consensus sequence (SSCS) calling

A custom Python script was written to generate single strand consensus sequences (SSCS) from the mapped BAM file in a stepwise manner. First, reads that had a mapping quality >20 and shared the same UMI tag, genomic coordinates and template length were grouped together to form 'UMI families' (Figure 4.7a). For the SNV/ indel panel, UMI families were discarded if they contained <3 reads. For the mCA/ chromosomal rearrangement panel, no minimum UMI family size was required. The reads in a UMI family were then compared at each sequence position, if their base quality score was >20, and a consensus nucleotide was called if there was at least 90% agreement between the reads (Figure 4.7b). If there was <90% agreement, then an 'N' was called as the consensus nucleotide at that position. This meant, for UMI families containing <10 reads, there had to be 100% agreement between the reads at a position for the base to not be called as 'N'. The resulting SSCS reads were written to an 'unmapped' SSCS BAM file. Reads that did not have a 99, 163, 147, 83 flag (i.e. were not 'mapped in proper pairs') were discarded, except for the generation of SSCS BAM files for FLT3-ITD calling and chromosomal translocation calling.

Duplex consensus sequence (DCS) calling

A custom Python script was written to generate duplex consensus sequences (DCS) from the 'unmapped' SSCS BAM file in a stepwise manner. First, SSCS reads corresponding to a pair of the initial DNA strands were identified and grouped together. The UMI tag associated with each read consists of two 3-nucleotide sequences and the UMI tag of the read from its partner strand is a transposition of this (Figure 4.7a). For example, if a 'read 1 forward' sequence had the UMI tag 'ATG-CAT', it would be grouped with a 'read 2 forward' sequence that had the UMI tag 'CAT-ATG', the same genomic coordinates and the same template length. The paired SSCS sequences were then compared at each sequence position and a consensus nucleotide was called if the bases matched. If one or both of the bases had a quality score <20, or if the bases did not match, then an 'N' was called as the consensus nucleotide at that position. The resulting DCS reads were written to an 'unmapped' DCS BAM file with 'forward strand' DCSs becoming 'read 1' reads and 'reverse strand' DCSs becoming 'read 2' reads.

Putative variant detection

For SSCS or DCS variant detection, the 'unmapped' SSCS or DCS BAM file was processed as follows: To identify any adapter sequences that may have been missed pre-consensus calling, Illumina adapter sequences were again marked using Picard *MarkIlluminaAdapters* and the adapter-marked SSCS or DCS BAM was converted to a fastq using Picard *SamToFastq*, which also clipped any remaining adapter sequences. The fastq was realigned to The Broad Institute's b37/hg19 reference genome[181] using BWA-MEM[180] to create a mapped BAM file. Picard *MergeBamAlignment* was used to transfer the BAM tags, containing SSCS or DCS calling metrics, from the unmapped adapter-marked BAM file to the mapped BAM file. Overlapping reads and 3 nucleotides from the end of each read were clipped and then sequences were realigned with GATK's Indel Realigner[181]. The aligned sequences were processed with SAMtools[182] mpileup using the parameters -BOa -Q0 -d 1,000,000 to ensure all the pileups were returned without any filtering. A custom Python script was then written to generate a VCF file which contained SNV and indel information as well as the variant depth and total read depth at every position in the panel. All positions with a variant depth >0 were annotated with ANNOVAR[185]. Indels were also called from the realigned BAM file using VarDictJava[183] and FLT3-ITD variants were called using Pindel[184].

For the SNV/ indel panel, variants were called from both SSCS and DCS files. The same depth accuracy was not required for the mCA, KMT2A PTD and chromosomal rearrangement panel and so only the SSCS was used for calling these.

4.3.2 Error-corrected sequencing metrics

In order to achieve as low a VAF limit of detection as possible, it is important to maximise the number of DCSs formed. Forming a DCS requires an SSCS to be generated from both DNA strands and so it is important that there are sufficient sequencing reads sharing the same UMI tag sequence to do this (i.e. sufficient UMI tag family size). The UMI tag family size is dictated by the number of input DNA fragments for PCR as well as the number of sequencing reads dedicated to the sample. If there is too much PCR input and/ or too few sequencing reads then the UMI tag family size will be too small and a SSCS (and therefore DCS) will not be called. Within a sample there is a distribution of UMI tag family sizes, due to the variation in PCR amplification efficiency between DNA fragments. A peak is typically seen at a UMI tag family size of 1, most likely due to sequencing errors in the UMI tags[137]. Excluding the peak at 1, an optimal peak family size of 6 is recommended to maximise the efficiency of duplex sequencing[137], striking a balance between optimal PCR input and sequencing depth/ cost. Only ∼8% of our pre-AML and control samples (SNV/ indel panel) had a peak family size $\geq$ 6, with nearly 75% of samples having a peak UMI tag family size of 2 (Figure 4.8a). Our UMI tag family size distributions were broad, however, with an average mean UMI tag family size of 8 and maximum UMI tag family size of ∼175 for both pre-AML and control samples (Figure 4.8b). This meant, even when we required a minimum UMI tag family size of 3 for an SSCS to be called, we still retained ∼92% of reads (Figure 4.8c), which represented ∼64% of UMI tag families (Figure 4.8d). Whilst it might be prudent to adjust PCR input and/ or sequencing coverage in the future to optimise sequencing costs, the mean SSCS:DCS ratio attained was 5-6 (Figure 4.8e), which is the same as in the original duplex sequencing protocol paper[137]. Across all samples, for the SNV/ indel panel, mean SSCS depth was ∼5500 and mean DCS depth was ∼1800. With ∼50ng DNA input (∼15,000 haploid genomes), this equates to an efficiency of ∼7.5% for SSCS and ∼6% for DCS.

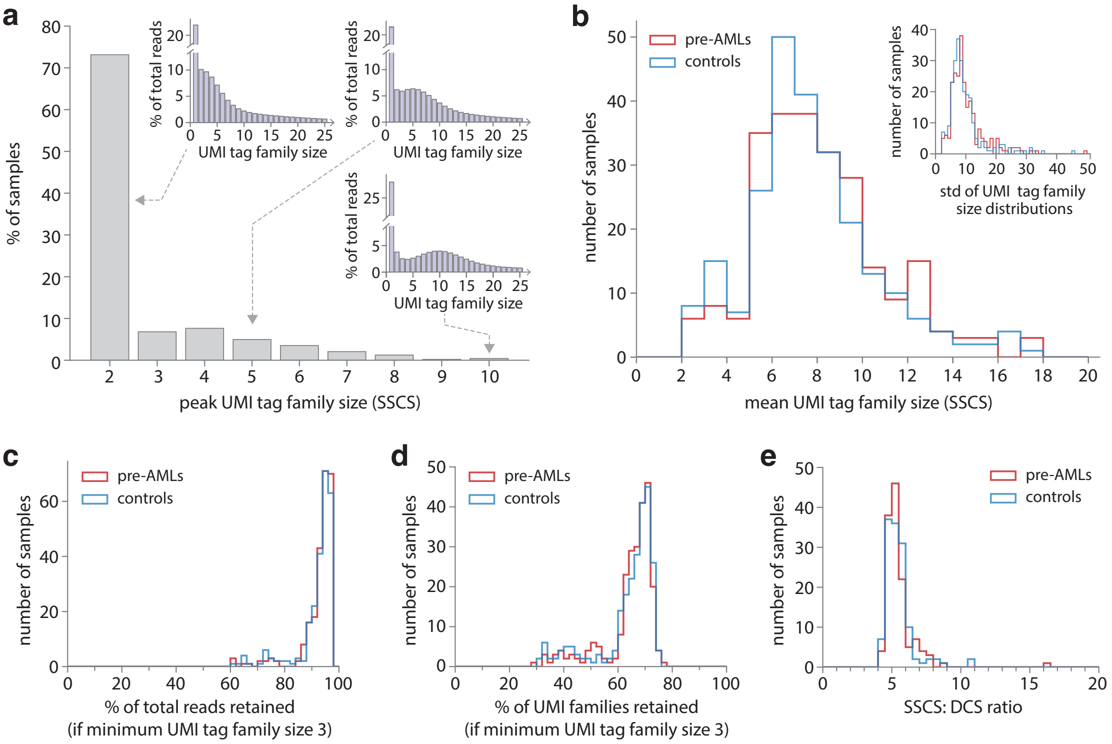


Figure 4.8 Error-corrected sequencing metrics (SNV/ indel panel). **a.** Distribution of peak UMI tag family sizes across all pre-AML and control samples. Insets show example UMI tag family size distributions with peaks of 2, 5 and 10. Example distributions have been capped at a UMI tag family size of 25. Average maximum UMI tag family size was 179 (range 60-1044) for pre-AML cases and 169 (range 57-875) for controls. **b.** Distribution of mean UMI tag family sizes across all pre-AML samples (red) and control samples (blue). Average mean UMI tag family size was 8 (range 2-17) for both pre-AML cases and controls. Inset shows the distribution of standard deviations (std) for the UMI tag family size distributions. **c.** Distribution of the proportion of total reads retained if those forming UMI tag family sizes of <3 are discarded. Mean proportion was 93% (range 61-98%) for pre-AML samples (red) and 92% (range 61-98%) for controls (blue). **d.** Distribution of the proportion of UMI tag families retained if those containing <3 reads are discarded. Mean proportion was 64% (range 29-77%) for pre-AML samples (red) and 64% (range 30-75%) for controls (blue). **e.** Distribution of ratio of SSCS to DCS reads across pre-AML samples and control samples. Mean SSCS:DCS ratio was 5.5 (range 4.3-16.2) for pre-AML cases (red) and 5.5 (range 4.4 to 10.9) for controls (blue).

4.4 Initial testing of the duplex sequencing approach

To validate the duplex sequencing approach it is important to assess the true positive rate (i.e. what proportion of true variants could be detected) as well as the false positive rate (i.e. how many false variants are detected). To do this a Myeloid DNA Reference Standard (Horizon Discovery Ltd.) was used, which contains 14 SNVs and 5 insertions/ deletions in gene regions targeted by our SNV/ indel panel at validated VAFs ranging from 5 - 70% (Appendix Table C.5). Serial dilutions were created (25%, 1% and 0.1%), by diluting the Myeloid DNA Reference Standard with a sample of peripheral blood-derived DNA from a 65 year old individual (hDNA). These serial dilutions, as well as 100% Myeloid DNA Reference Standard and 100% hDNA were sequenced in replicate on the same sequencing lane as the pre-AML and control samples (i.e. NovaSeq S4 XP workflow, with $\sim$ 68 million reads per sample). The sequencing files were then processed using the computational workflow described in Section 4.3.1. Variants were filtered out if the total depth at the position was <500, the mean variant position in the read was <8 or the variant's MAF in ExAC[186] was >1%.

There was good concordance between replicate samples (Figure 4.9a) and 88% of the validated variants with expected VAFs >0.1% were successfully called from the DCS (Appendix Table C.6). Only 41% of validated variants with VAFs of 0.01-0.1% could be detected, which reduced to 15% for those with VAFs of 0.001-0.1%. Our reliable VAF limit of detection, is therefore likely close to $\sim$ 0.1%, which is consistent with the mean DCS depth of $\sim$1800. We found however, that we were calling far more variants at frequencies <0.5% than we expected (white data points Figure 4.9a), in both the Myeloid Reference Standard DNA and the hDNA, the majority of which were detectable in only one of the replicate samples. Analysis of the serial dilutions revealed many of these 'variants' at <0.5% VAF to be errors, as they were found at the same frequencies in the 100% Myeloid Reference Standard as the dilutions (cluster of variants lying at frequencies <1% around the grey dotted line, (Figure 4.9b). Further analysis of specific base changes revealed C>T errors were particularly common (Figure 4.9c) in both the Myeloid Reference Standard as well as the hDNA sample.

How many variants do we expect to see?

From our analysis of blood sequencing data from $\sim$50,000 individuals (Chapter 2) we know that the distribution of clone sizes is consistent with a simple branching process model of haematopoietic stem cell (HSC) divisions[93], such that the probability density as a function of VAF (f) is given by:

$$\rho(f) = \frac{2N\tau\mu}{f(1-2f)} e^{-\frac{f}{\phi(1-2f)}} \qquad \text{where} \qquad \phi = \frac{e^{st}-1}{2N\tau s} \tag{4.1}$$

N = total number of HSCs

τ = time in years between successive symmetric cell differentiation divisions

μ = mutation rate per year, t = age in years, s = fitness effect of variant (% per year)

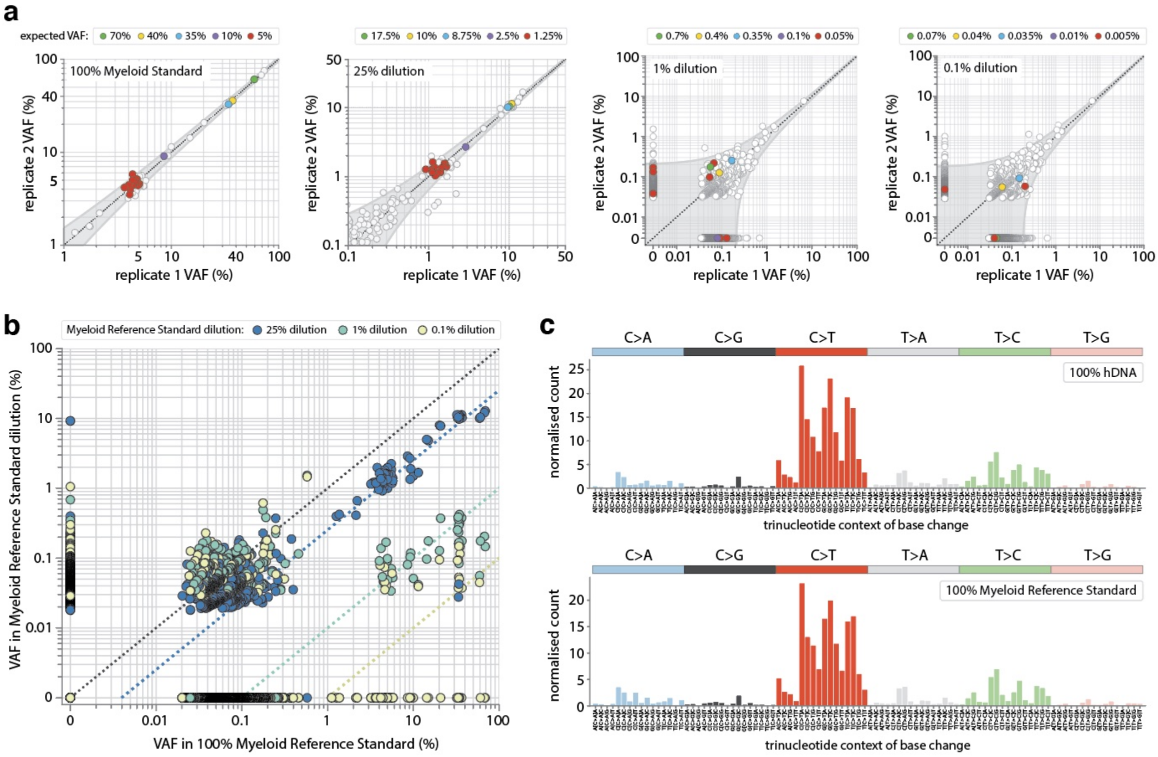


Figure 4.9 DCS variant calling using serial dilutions of Myeloid Reference Standard DNA. **a.** Concordance between replicates for serial dilutions of Myeloid Reference Standard DNA. Validated variants in the reference standard are shown in colour, according to their expected VAF. **b.** Correlation between variant VAFs in the undiluted Myeloid Reference Standard and the VAF in the diluted sample. Variants present in hDNA are not shown. **c.** Mutational signatures for variants called in the undiluted hDNA sample (top) and the undiluted Myeloid Reference Standard sample (bottom).

Calculating the number of variants we expect in the manufactured Myeloid Reference Standard is difficult, but we can use this equation to calculate a rough estimate of the number of variants we expect to see in the hDNA sample, which was collected from a 65 year old (t) individual. Single-cell derived sequencing work by Lee-Six[48] provides us with HSC mutation rate estimates of $\sim 2.7 \times 10^{-9}$ /bp /year, which, multiplied by the size of our SNV/ indel panel ($\sim$ 58 kb), gives a mutation rate (μ) estimate of $\sim 1.6 \times 10^{-4}$ per year. Our work has provided estimates for $N\tau$ of $\sim 100,000$ as well as the distribution of fitness effects (s) of variants across 10 of the most commonly mutated clonal haematopoiesis genes (Chapter 2)[93]. Using these parameters and integrating our expected density of clones (eq. 4.1) over the distribution of fitness effects and then over a range of VAFs (e.g. from 0.1% to 50%), allows us to estimate the number of variants we should expect to observe above a certain VAF in an individual of a particular age (Figure 4.10a). Bearing in mind that using this distribution of fitness effects likely provides an over-estimate for the number of variants we would expect to see across all genes, we can roughly estimate that we should expect to see $\sim$ 8 variants between 0.1% and 50% VAF in a 65 year old individual, of which $\sim$ 5 will be at VAFs < 0.5%. In the hDNA sample, however, we see 252 putative variants between 0.1% and 50% VAF, of which 240 are at VAFs <0.5% (Figure 4.10b).

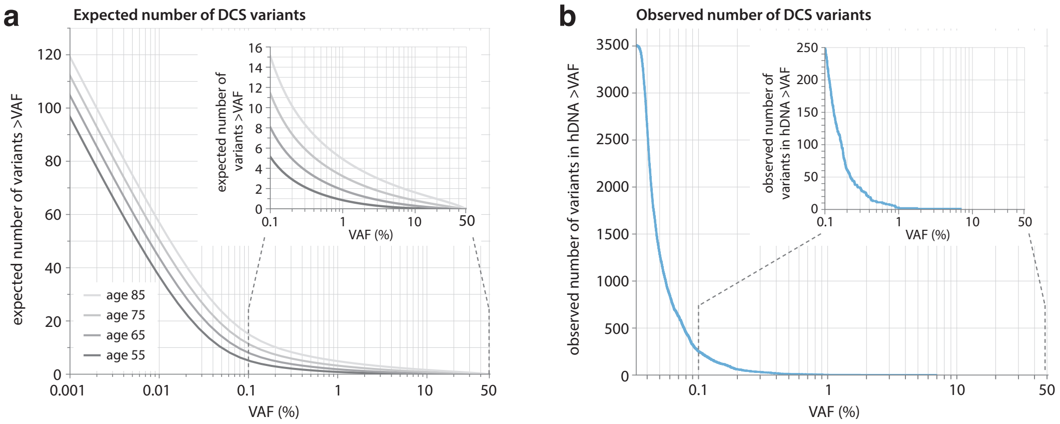


Figure 4.10 Expected and observed number of DCS variants. **a.** Expected number of variants detectable above a given VAF, using the custom SNV/ indel panel, for individuals aged 55, 65, 75 and 85 years. **b.** Number of DCS 'variants' observed above a given VAF, using the custom SNV/ indel panel, in hDNA replicate 1 (from a 65 year old individual).

The theoretical error rate of duplex sequencing is often quoted as $< 10^{-9}$ error per bp, which is simply the probability of two complementary errors occurring at the same nucleotide position on both DNA strands, either spontaneously or during the 1st PCR cycle[140,187]. In reality, however, library preparation artefacts e.g. due to sonication, end-repair and mapping errors[14,141,188] have all been shown to slip through the duplex error correction and likely explains why we see far more 'variants' than we expect. Therefore, in order to reliably call 'true variants' <0.5% VAF we developed an *in silico* noise correction method for SNV calling, which involved developing a null model for the errors, so that inconsistencies from the model would be identified as 'true variants'.

4.5 Developing an *in silico* noise correction method for SNV calling

To develop an *in silico* noise correction method for SNV calling we used sequencing data from 40 samples (20 pre-AMLs and 20 controls), which had all been sequenced on the same Illumina NovoSeq S4 lane. The sequencing data was processed using the computational workflow described in Section 4.3.1 and the DCS output files, containing variant depth and total depth at all positions in the SNV/indel panel, were used for error model development.

4.5.1 Trinucleotide-context specific error-model

We first considered whether error rates were trinucleotide-context specific (e.g. C[G>A]T and G[T>A]T would have separate defined error rates). Our null hypothesis was that each trinucleotide context had a specific error-rate (ε) and the number of variant reads (k) observed at that context would be consistent with binomial sampling at that context's error rate (Figure 4.11, eq. 4.2):

$$P(k) = \binom{N}{k} \varepsilon^{k}(1-\varepsilon)^{N-k} \tag{4.2}$$

where N = number of positions with a particular trinucleotide-context

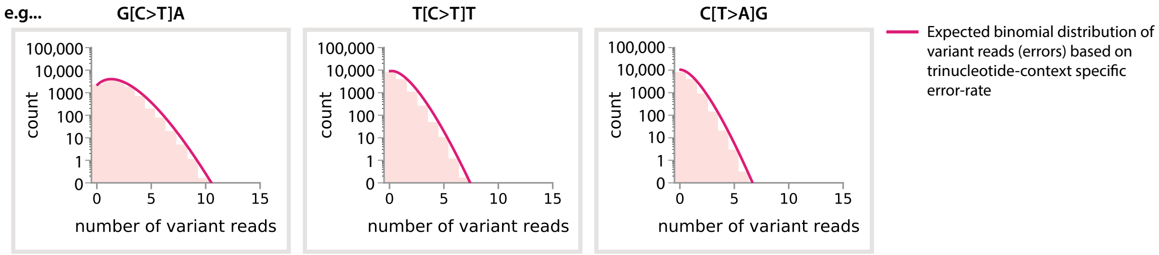


Figure 4.11 Schematic of trinucleotide context specific error rates.

The problem with comparing the distribution of variant reads to an expected binomial distribution, however, is that there are ∼175,000 possible base change positions (3 at each position, e.g. C>T, C>A, C>G) across the panel and calculating a binomial distribution becomes cumbersome for large N. Because we have both large N and small ε, the binomial distribution can be well-approximated by a Poisson distribution. To check the null hypothesis, therefore, the distribution of variant reads (k) for each of the different trinucleotide contexts was compared to an expected Poisson distribution:

$$P(k) = N\frac{\lambda^k e^{-k}}{k!} \tag{4.3}$$

$$\text{where} \quad \lambda = D\varepsilon, \qquad D = \text{mean depth} \qquad \text{and} \qquad \varepsilon = \sum \text{variant reads} \Big/ \sum \text{depths}$$

Because the distribution of variant reads will be affected by the distribution of depths (i.e. more variant reads would be expected in a region with greater depth), we take this into account in the expected Poisson distribution by summing over the expected Poisson distributions for small intervals of depth (width 100) with ‘mean depth’(D) and N (number of positions with depth in that interval) calculated separately for each interval (eq. 4.4). The ε (error rate) was calculated across all depths and was therefore constant across intervals:

$$P(k) = \sum_{i=1}^{C} N_i \frac{(D\varepsilon_i)^k}{k!} e^{-D\varepsilon_i} \tag{4.4}$$

where N and D are interval specific and C=number of depth intervals

When calculating the error rate ε for each position, the assumption is that all the variants seen at that position are errors. To try to exclude true variants, which would skew the estimated error rate (particularly SNPs at 50% VAF), samples whose VAF at a particular position was >1% were not included in the calculation of the error-rate at that position.

As can be seen from Figures 4.12, 4.13 and 4.14, although the distribution of variant reads at some trinucleotide contexts is consistent with our null hypothesis (e.g. C[C>G]A, C[T>G]A), the distribution at the majority of contexts is far broader than expected. It seems unlikely, therefore, that the error-rates are consistently trinucleotide-context dependent.

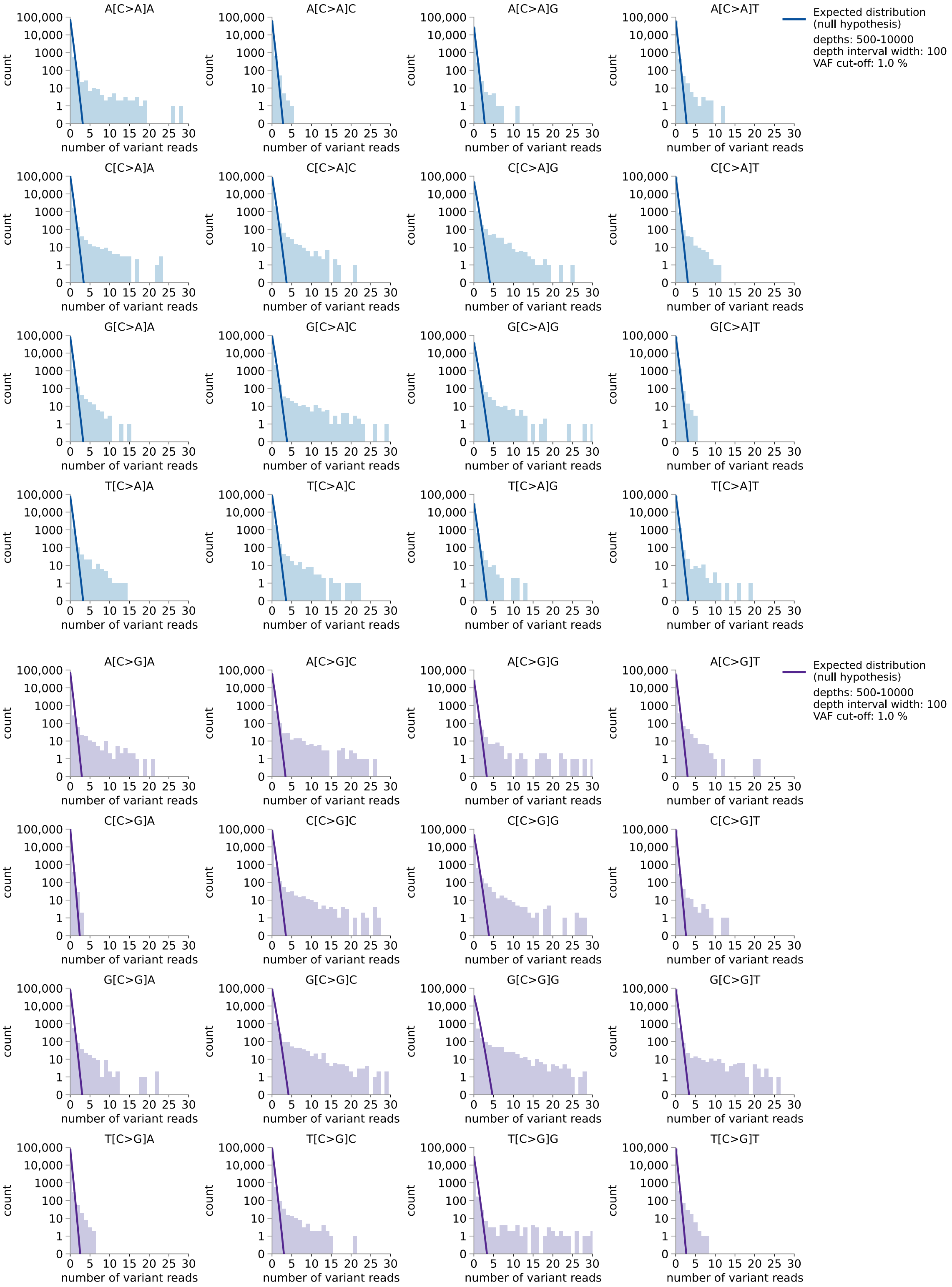


Figure 4.12 **Distribution of DCS variant reads by trinucleotide context: contexts with C>A and C>G base change**. The darker solid line represents the expected Poisson distribution of variant reads (i.e. if the number of variant reads (errors) at each trinucleotide context was consistent with binomial sampling). Samples where the total depth at a position was <500 or >10000 were not included in the calculation of the position's error-rate, nor were their variant reads included in the histogram.

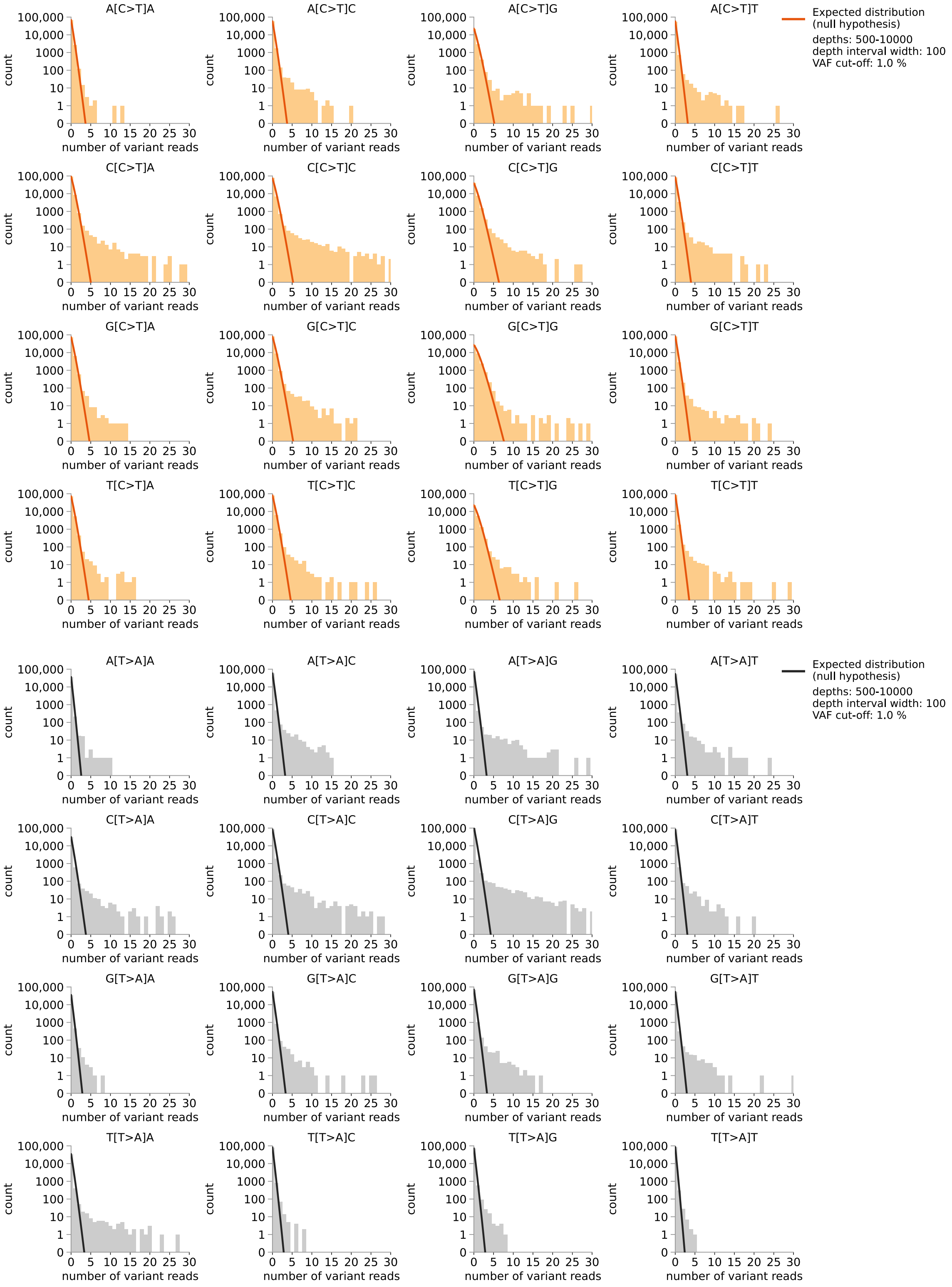


Figure 4.13 **Distribution of DCS variant reads by trinucleotide context: contexts with C>T and T>A base change**. The darker solid line represents the expected Poisson distribution of variant reads (i.e. if the number of variant reads (errors) at each trinucleotide context was consistent with binomial sampling). Samples where the total depth at a position was <500 or >10000 were not included in the calculation of the position's error-rate, nor were their variant reads included in the histogram.

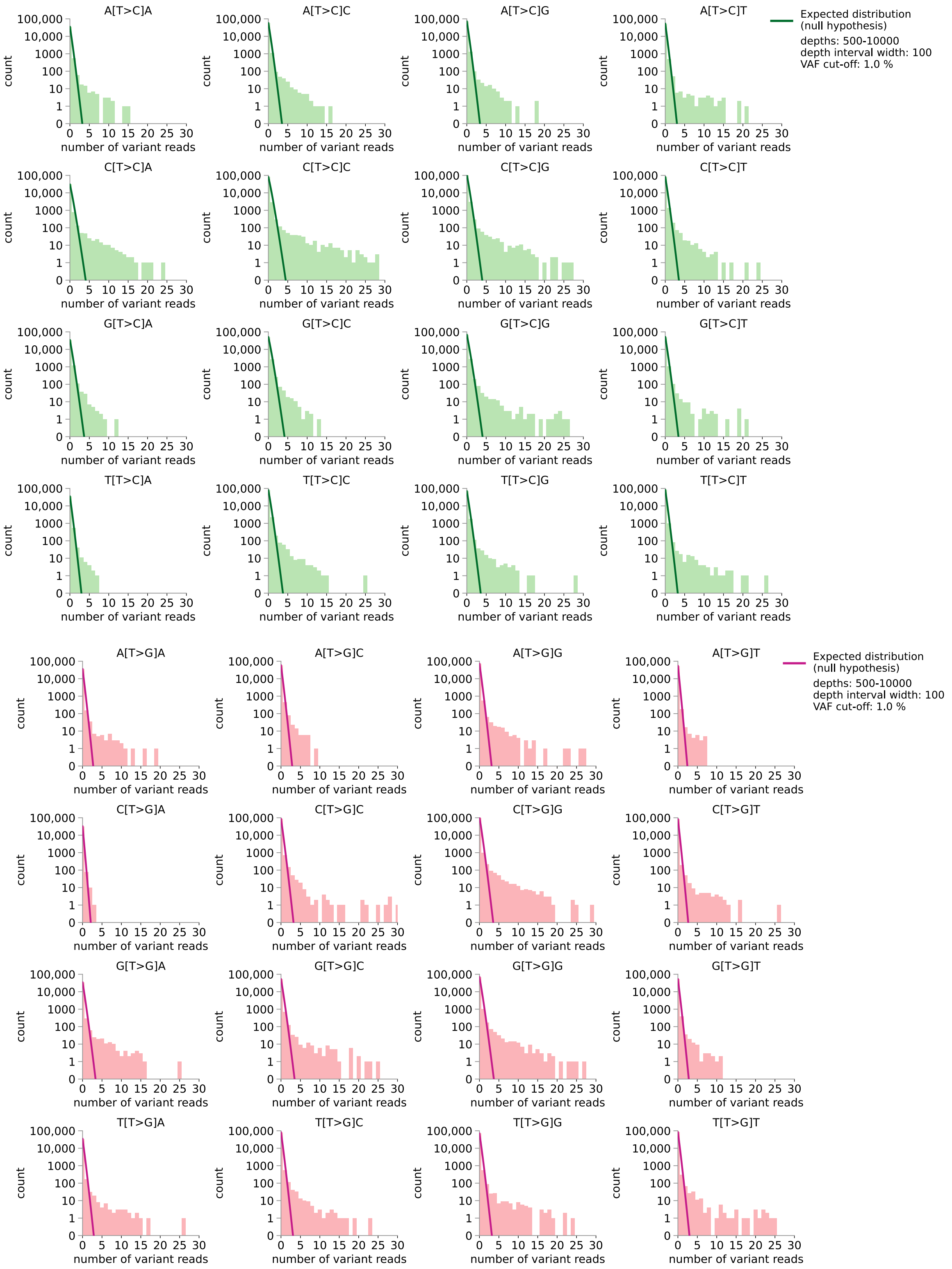


Figure 4.14 Distribution of DCS variant reads by trinucleotide context: contexts with T>C and T>G base change. The darker solid line represents the expected Poisson distribution of variant reads (i.e. if the number of variant reads (errors) at each trinucleotide context was consistent with binomial sampling). Samples where the total depth at a position was <500 or >10000 were not included in the calculation of the position's error-rate, nor were their variant reads included in the histogram.

4.5.2 Position-specific error model

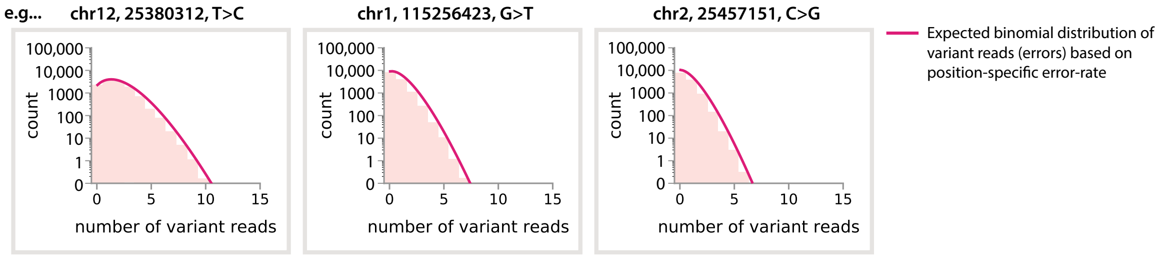


Figure 4.15 Schematic of position base-change specific error rates.

We next considered whether error-rates were position specific (e.g. ‘chr 12, pos 25380312, T>C’ and ‘chr1, pos 115256423, G>T’ would have separate defined error rates). Our null hypothesis was that each base-change at a position in the panel had a specific error rate (ε) and the number of variant reads (k) observed in a sample at that position would be consistent with binomial sampling at that specific error rate (Figure 4.15). When considering many samples, the variation in total read depths between samples makes the distribution of variant reads difficult to interpret and so the distribution of ‘VAFs’ (variant reads/ total depth) across samples at a position base-change was visualised instead. We started by assuming that all ‘variants’ with the same position base-change could be real if their DCS ‘VAF’ was >1% and then calculated the position’s mean error rate, for that base change, from the ‘VAF’s of all the other samples. Looking at some example position-specific base-changes from across the panel (Figure 4.16a), we can see that whilst the majority appear consistent with a binomial position specific

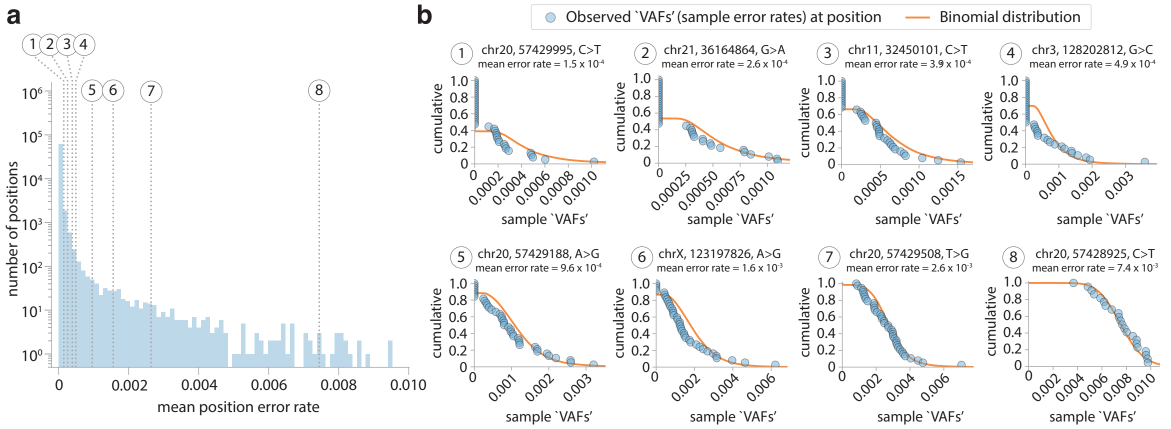


Figure 4.16 Distribution of sample DCS ‘VAFs’ (error-rates) at example positions. a. Distribution of mean position error rates across the SNV/indel panel. Samples with a position depth <500 were excluded from the error rate calculation and positions that had 0 variant reads across all samples at the position base-change are not shown in the distribution. Example positions across the distribution were chosen (numbered) to examine in more detail. **b.** 8 positions, that had at least 1 variant read detected in at least 16 samples with minimum sample depth of 500 and ‘VAF’ <1% were chosen from the distribution of mean position base-change error rates across the SNV/indel panel (in a.). Each sample’s ‘VAF’ (error rate) for the base-change at that position is shown in blue (cumulative distribution). If no variant reads were detected for that base-change in the sample then the error rate was recorded as 0. The orange line represents a binomial distribution, summed across all samples at the position where n = sample depth and p = mean position base-change error rate.

error model there are some that appear slightly over-dispersed (Figure 4.16b). The null hypothesis was therefore adapted to consider a beta-binomial position base-change specific error model, which allows for a wider variation in the sample errors at a given position base-change, controlled by a parameter δ:

$$P(k) = \binom{D}{k} \frac{\mathrm{B}(k+\alpha, N-k+\beta)}{\mathrm{B}(\alpha,\beta)} \tag{4.5}$$

$$\text{where} \quad \mathrm{B} \text{ is the Beta function,} \quad D = \text{total depth,} \quad \alpha = \delta^{-1}, \quad \beta = \frac{1-\varepsilon}{\delta\varepsilon}$$

$$\varepsilon = \text{position-specific error rate,} \quad k = \text{number of variant reads}$$

The binomial distribution naturally emerges from the beta-binomial distribution in the limit of $\delta\varepsilon \ll 1$ (Appendix C.1.2) and so an advantage of a beta-binomial model is that it is also able to capture the behaviour of positions that are consistent with a simple binomial model.

Inferring ε and δ parameters for identification of real variants

A custom Python script was written to infer the error rate (ε) and dispersion (δ) values for each possible single nucleotide base-change at each position in the panel (hereafter referred to as simply 'position') and samples were called as 'real' variants if their variant read count was inconsistent with this distribution of errors. First, samples with a VAF >10% at the position were automatically called as real and excluded from the subsequent analysis. Then, the remaining samples at the position were used to fit either a beta-binomial distribution or a binomial distribution, depending on the number of samples at the position that had $\geq$ 1 variant reads.

If there were $\leq$ 3 samples remaining with $\geq$ 1 variant reads, there was insufficient data to fit a beta-binomial distribution and so the distribution of errors was assumed to be binomially distributed with $\varepsilon = \sum$variant reads/$\sum$depth. The binomial *p*-values for all the samples with $\geq$1 variant reads were calculated and a sample was called as real if its *p*-value was less than a chosen *p*-value threshold.

If there were > 3 samples with $\geq$ 1 variants reads, a beta-binomial distribution was fitted to all the samples at the position. First, ε was estimated as $\sum$variant reads/$\sum$depth across all the samples to be included in the fit and δ was estimated using the method-of-moments estimator (Box 4.5.2a). These δ and ε estimates were then used to initialise a maximum likelihood approach which re-inferred δ and ε by minimising the negative log likelihood of the model (Box 4.5.2b and Appendix C.1.1). If the method-of-moments estimate for δ was <1, suggesting the data was either undispersed or underdispersed relative to the binomial, then δ was initialised as 10^{-4} in the maximum-likelihood approach. A lower bound for δ was set to limit δ/ε to $>10^{-8}$ as the distribution had already become binomial at $\delta\varepsilon < 10^{-6}$ and numerical issues occurred below this level due to large α and β values (Appendix C.1.2). The beta-binomial *p*-values for all the samples with $\geq$1 variant reads were calculated and a sample was called as a real variant if it's *p*-value was less than a chosen *p*-value threshold.

Box 4.5.2a: Method of moments estimator for δ

The method of moments estimator for δ can be calculated by noting the 1st and 2nd moments of the beta-binomial...

$\mu_1 = \frac{D\alpha}{\alpha+\beta}$ (1st moment)

$\mu_2 = \frac{D\alpha[D(1+\alpha)+\beta]}{(\alpha+\beta)(1+\alpha+\beta)}$ (2nd moment)

where D = mean total depth, α = δ^{-1}, β = $\frac{1-\varepsilon}{\delta\varepsilon}$ and ε = position error rate

Setting these raw moments equal to the 1st and 2nd raw sample moments...

$\hat{\mu_1} := m_1 = \frac{1}{S}\sum_{i=1}^{S} k_i$

$\hat{\mu_2} := m_2 = \frac{1}{S}\sum_{i=1}^{S} k_i^2$

where S = number of samples and k = number of variant reads in a sample

and solving for α and β we get...

$\hat{\alpha} = \frac{Dm_1 - m_2}{D(\frac{m_2}{m_1} - m_1 - 1) + m_1}$

$\hat{\beta} = \frac{(D-m_1)(D-\frac{m_2}{m_1})}{D(\frac{m_2}{m_1} - m_1 - 1) + m_1}$

Given δ = $1/\alpha$, we can use this estimate for α to estimate δ.

Box 4.5.2b: Maximum likelihood approach for inferring ε and δ parameters

For each possible base change at a position, testing different position error rates (ε) and beta-binomial dispersion values (δ)...

1. For each sample, calculate the beta-binomial likelihood of measuring that number of variant reads for that base change at that position (k), given the sample depth (D), position specific base change error rate (ε) and dispersion value (δ):

   $f(k \mid D, \varepsilon, \delta) = \binom{D}{k}\frac{\mathrm{B}(k+\alpha, D-k+\beta)}{\mathrm{B}(\alpha,\beta)}$ where B is the Beta function, α = δ^{-1} and β = $\frac{1-\varepsilon}{\delta\varepsilon}$

2. Calculate the likelihood of the model by multiplying across all the sample Beta-binomial likelihoods:

   $$L(\varepsilon, \delta) = \prod_{i=1}^{S}(f(k \mid D, \varepsilon, \delta))$$ where S = total number of samples at that position

   This creates very small numbers and so an alternative is to sum across all the sample log(Beta-binomial likelihoods):

   $$L(\varepsilon, \delta) = \sum_{i=1}^{S}\log(f(k \mid D, \varepsilon, \delta)$$ where S = total number of samples at that position

3. Find the values of ε and δ that maximise the likelihood of the model (or minimize the negative likelihood of the model).

An iterative approach at potential 'hotspot' sites

When fitting a binomial or beta-binomial error distribution to a position only once, a problem arises if more than one sample at the position has a 'real variant'. The problem is that the lower VAF variants will not be called as they will be fitted as 'errors' within a falsely over-dispersed beta-binomial distribution or a binomial distribution with a falsely elevated error rate. We therefore chose to use an iterative approach for positions observed in haematopoietic and lymphoid tissues in COSMIC v92[166] ($\sim 2\%$ of sites across our custom panel), where we might expect to see more than one sample with a real variant in our pre-AML cohort.

For positions with ≤ 3 samples with ≥ 1 variant reads, the iterative approach involved fitting a binomial distribution to all the samples, calling variants as real if their *p*-value was less than a chosen threshold, removing the real variants and then fitting the binomial again. This was then continued at the position until no further real variants were called.

For positions with > 3 samples with ≥ 1 variants reads, the iterative approach was similar except at each iteration the sample with the highest VAF was excluded from the beta-binomial fit. This is because our prior probability of there being at least one real variant at these potential 'hotspot' positions is higher and if it were at high VAF it would result in a falsely over-dispersed beta-binomial distribution being fitted and real variants being missed. Beta-binomial *p*-values were then calculated for all samples, including the highest VAF sample, and variants were called as real if their *p*-value was less than a chosen threshold. This iterative approach (excluding the highest VAF sample from the beta-binomial fit each time) was then continued at the position until no further real variants were called. The highest VAF sample was not excluded when fitting beta-binomial distributions at other positions because we found, when exploring different approaches, that this resulted in an increased false positive rate (Appendix C.1.3). Similarly, performing an iterative approach also resulted in an elevated false positive rate (Appendix C.1.3). We therefore chose to use these two adaptations only at sites where there was an increased prior probability of a variant being real.

An iterative approach for longitudinal samples

In the majority of cases, different timepoint samples from the same individual were sequenced on different flow-cells. When there was more than one timepoint sample from the same individual on the same flow-cell, an iterative approach was used for positions at which one of those samples had been called as a real variant. This is because, for these positions, there is a much higher chance that one of the other timepoints also contained the same variant, but at a different VAF. Without an iterative approach, it is likely the sample with the lower VAF would not be called as real. The iterative approach at the position was continued until no further samples from the individual with multi-timepoint samples were called as real.

Choosing an appropriate p-value threshold for calling real variants

To calculate an appropriate p-value threshold, we examined the false discovery rate (FDR) as a function of p-value threshold. Correctly estimating the FDR relies on the assumption that the underlying null model is correct, such that the number of false positives reliably increases as the p-value threshold increases. To check this, a custom Python script was written to simulate 40 samples across $\sim$575 positions, with the number of variant reads in each sample chosen from a beta-binomial distribution with a position specific ε and δ (i.e. all 'variants' were errors). Each position had a different ε and δ, covering regularly distributed combinations of ε between 10^{-3} and 10^{-1} and δ between 10^{-2} and 1. A beta-binomial distribution (or binomial if ≤ 3 samples with ≥ 1 variant reads) was fitted at each position (non-iteratively), p-values for each sample calculated and 'real' variants called if the sample's beta-binomial (or binomial) p-value was less than the p-value being tested. The number of variants called across all positions was plotted as a function of p-value threshold (Figure 4.17a). This showed that the false positive rate was indeed consistent with what we would expect for most p-values, although there was a slight decrease at very low p-values. This may be a reflection of the fitted beta-binomial distribution slightly over-estimating the dispersion (δ) parameter and provides further support for our decision to exclude the highest VAF sample when fitting the beta-binomial distribution at 'hotspot' sites.

When we tested a range of different p-value thresholds on our real samples (38 samples from 1 NovaSeq 6000 S4 lane), the false positive rate appeared consistently much lower than expected (Figure 4.17c). Initial analysis of the position error-rates had shown that most position error rates were very low, at $< 10^{-4}$ (Figure 4.16), and so we repeated the simulated false positive analysis (as

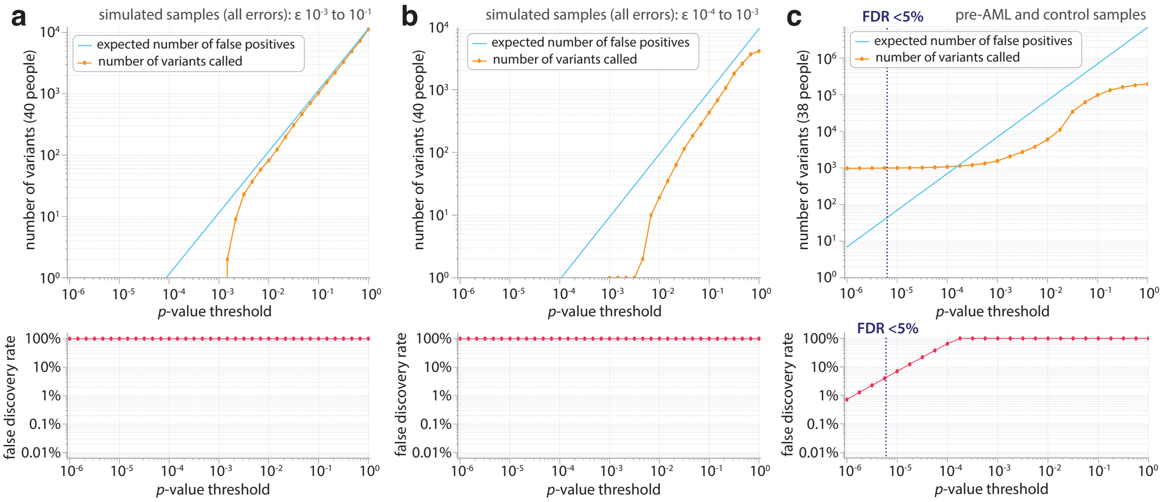


Figure 4.17 Choosing a p-value threshold for calling real variants. The point at which the read data (orange line) starts to deviate from the blue line, represents the p-value threshold where there are equal numbers of false positives and true positives. If we lower the p-value threshold further, the proportion of true positives increases, but we also increase the false negative rate.

described above), but for positions where the error rate was much lower (ε: 10^{-4} to 10^{-3}, δ: 10^{-3} to 1). Interestingly, we saw a similar pattern to the real samples, where the false positive rate was consistently lower than expected (Figure 4.17b). We reasoned that this may be a consequence of the analysis being across only 40 samples, which means at very low ε there aren't any samples that have ≥ 1 variant reads and so no false positive variants are called. Therefore, to chose a *p*-value threshold, we chose a *p*-value that gave an FDR of $\sim 5\%$ (*p*-value: 6×10^{-6}), with the understanding that the actual FDR will likely be even lower than this (i.e. FDR <5%) and that it is difficult to accurately estimate the FDR for this number of samples when the average error-rate is so low.

The prior probability of a variant being real is higher if it has been called as real in a sample collected at a later timepoint. We therefore used a *p*-value threshold of 0.1 to call a variant as real if it had already been called as real with a *p*-value threshold of 6×10^{-6} in the final timepoint sample.

Because different sequencing runs could introduce potential sequencing run-related noise, the error model was applied to samples sequenced on the same sequencing lane (i.e. ~ 40 samples). An alternative approach would be to allow greater dispersion.

Position specific distributions of errors

Once the 'real' variants had been called at each position, a final beta-binomial or binomial distribution was fitted to the remaining samples at the position to ascertain each position's distribution of errors. Across 20 pre-AML cases and 20 controls, real variants were called at 197 positions out of 174,366 positions (Figure 4.18). Of the positions where no variants were called, $\sim$60% contained only samples with 0 variant reads (Figure 4.18: A positions) and $\sim 30\%$ of positions were fitted using a binomial distribution because there were ≤ 3 samples with ≥ 1 variant reads (Figure 4.18: B, C positions). The fitted beta-binomial distribution was consistent with a binomial distribution in $\sim$5% of positions (Figure 4.18: D-F positions) and a beta-binomial distribution in $\sim 3\%$ of positions (Figure 4.18: G-L positions). Of the positions where 1 or more real variants were called (197 positions), $\sim$40% had only samples with 0 variant reads remaining and $\sim 35\%$ were fitted using a binomial distribution because there were ≤ 3 samples remaining with ≥ 1 variant reads after the real variants were called. The fitted beta-binomial distribution was consistent with a binomial distribution in $\sim 11\%$ of positions (Figure 4.18: P, Q positions) and was consistent with a beta-binomial distribution in $\sim 10\%$ of positions (Figure 4.18: R-T positions). Overall, across all positions, 61% contained only samples with 0 variant reads, 36% were consistent with a binomial distribution of errors, and 3% were consistent with a beta-binomial distribution of errors. Looking at example positions from across the range of error distributions, we can see that the model appears to be performing well, appropriately fitting a beta-binomial distribution to positions whose distribution of errors appears over-dispersed relative to a binomial (Figure 4.18 subplots).

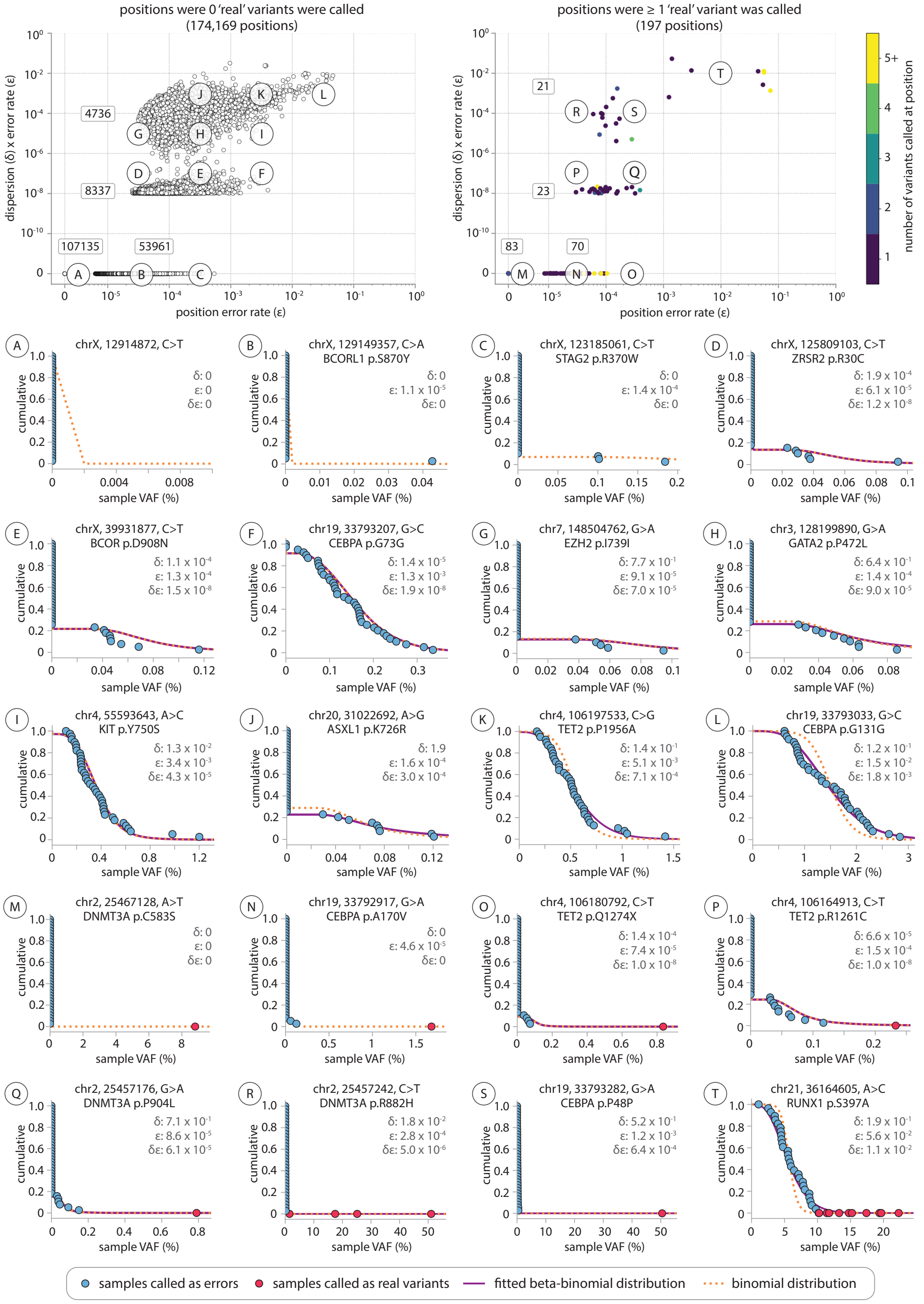


Figure 4.18 Position specific error distributions across 20 pre-AML and 20 control UKCTOCS samples. The relationship between the final position error rate (ε) and $\delta\varepsilon$ is shown for positions where no variants were called as real (top left plot) and positions where 1 or more variants were called as real (top right plot). The number of positions in each area is shown on the plot. Example positions from different δ and $\delta\varepsilon$ regions on the plot (indicated by letters A to T) are shown below. The final fitted error distribution is plotted as a purple solid line (beta-binomial) or orange dashed line (binomial). The binomial distribution is also shown for beta-binomially distributed positions for comparison. Samples called as real variants are shown in red. Samples called as errors are shown in blue.

Overall panel error rate

The overall error rate across the panel can be calculated as: ∑ variant reads/ ∑ depth for the samples that were not called as real variants, and was 2.6×10^{-5} across our initial test of 20 pre-AML cases and 20 controls (one sequencing lane). This means, on average, we would be unable to call variants at frequencies lower than this (i.e. VAF detection limit > 0.026%). Categorising the positions by their base change, we can see that C>T and G>A variants tend to have an error rate ∼5x higher than other base changes, with error rates as high as $\sim 10^{-2}$ (Figure 4.19). The distribution of position error rates is quite broad, with some positions having error-rates as high as nearly 10^{-1}, meaning there are some positions where we would be unlikely to call real variants unless they had a VAF > 10%.

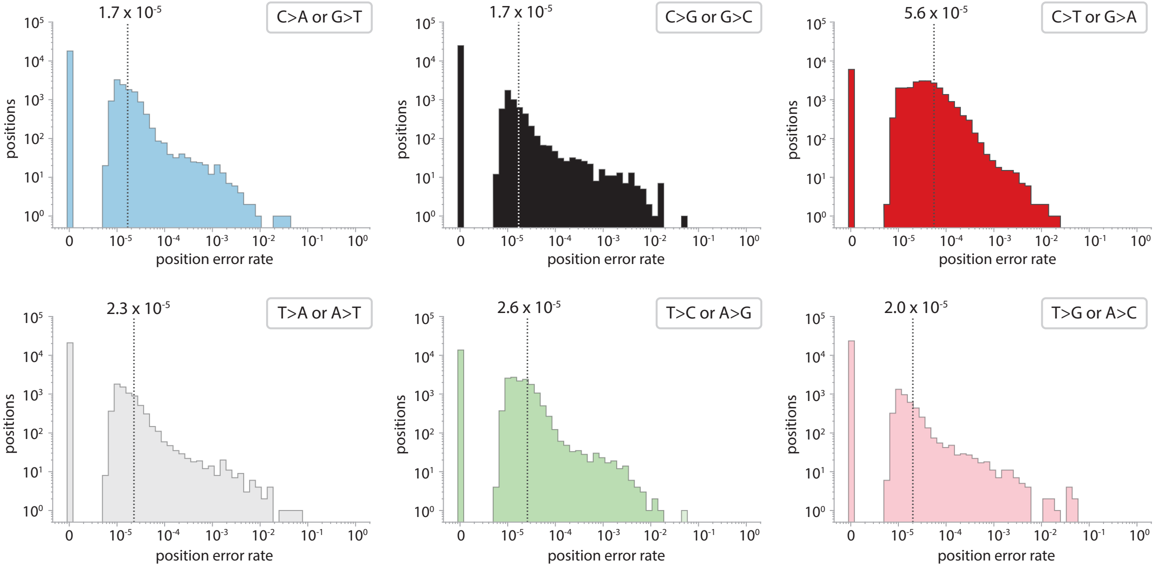


Figure 4.19 Distribution of final position-specific error rates, grouped by base change. The overall base-change error-rate is shown (dashed line) and was calculated as ∑ variant reads/ ∑ depth for the samples that were not called as 'real variants' at all the positions with the particular base change.

4.5.3 Post-processing of variant calls

The error model was applied to all the samples sequenced on the same sequencing lane (∼ 40 samples) and then additional post-processing filters were applied in order to filter out any additional errors that may have inadvertently been called as real. Variants were excluded if they were detected in only 1 variant read or if the total read depth at the position was less than 2 standard deviations below the mean read depth for that sample. Variants were also excluded if they were seen in ≥5% of people (e.g. Figure 4.18 position T), unless the variant had been observed in haematopoietic and lymphoid tissues in COSMIC v92[166].

Having longitudinal samples provided us with additional power for identifying errors, as we were able to use information from variants that were called, or not called, at multiple timepoints. Variants were excluded if their p-value was not $< 10^{-10}$ at ≥ 1 timepoint, unless the variant was present at >10% VAF at any timepoint. There were variants that were not detected at some timepoints, but were detected at the timepoints before and after. The expected growth rate was calculated from the VAFs at the two detected timepoints (VAF_1 at t_1 and VAF_3 at t_3) to determine how many reads (k) the variant should have been expected to be seen in at the missing timepoint (t_2), taking in to account the read depth of the missing sample:

$$\text{growth rate } (s) = \frac{\log\left(\frac{\text{VAF}_3}{\text{VAF}_1}\right)}{t_3 - t_1} \tag{4.6}$$

where t_1 is the age in years at the time of the VAF_1 sample and t_3 is the age in years at the time of the VAF_3 sample

$$\text{expected variant reads } (k) \approx \text{VAF}_1 \times e^{s(t_2 - t_1)} \times \text{sample read depth} \tag{4.7}$$

where t_2 is the age in years at the missing timepoint

The chance of not detecting the variant at t_2 was then calculated:

$$\text{chance of not detecting variant} = e^{-k} \tag{4.8}$$

If the chance of not detecting the variant was <10%, then the variants from the preceding timepoints were excluded, as it was deemed likely to also be an error. If a detected variant was missing from more than 1/3 of the timepoints, the chance of not detecting the variant was calculated at all the missing timepoints and if the combined chance of missing all of them was <10%, then the variant was excluded from all timepoints.

4.5.4 Testing the position-specific method on UKCTOCS samples

Before the *in silico* noise correction method was applied to our initial 20 pre-AML and 20 control UKCTOCS final timepoint samples, the number of variants we detected at <1% VAF was ∼100-1000-fold higher than we expected (Figure 4.20a). After applying the *in silico* noise correction method, however, the number of real variants we called was reassuringly in close agreement with that observed in other studies (Figure 4.20b). The density of variants we detect starts to decline below 0.3% VAF, suggesting that 0.3% VAF is our reliable limit of detection for the longitudinal final timepoint samples. Whilst it might seem disappointing that our limit of detection is not as low as the Young 2016 and 2019 studies, these studies used 250 ng input DNA for their SSCS error-corrected sequencing approach, which is $\sim 5 \times$ more than we used. Also, when calling variants from earlier timepoint samples, we can use the information from these final timepoint samples to increase our prior for lower VAF variants being real. Our limit of detection across all the UKCTOCS samples will therefore actually be much lower and at some positions we should be able to call variants down to almost single molecule resolution.

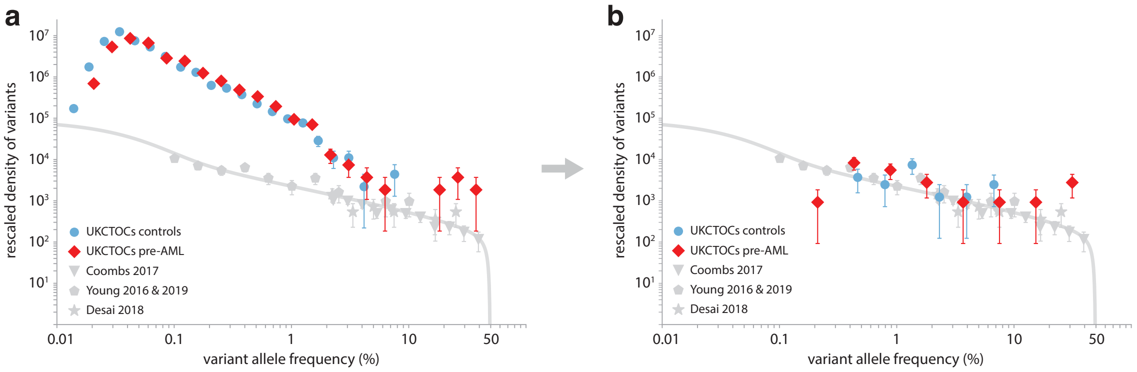


Figure 4.20 Effect of *in silico* noise reduction method on reducing the number of variant calls for the top 10 mutated CH genes (DNMT3A, TET2, ASXL1, TP53, SRSF2, SF3B1, JAK2, IDH2, KRAS, CBL). The densities of variants called across these genes in 4 studies (Coombs 2017[75], Desai 2018[76] and Young 2016[7] & 2019[39]) are shown for comparison (from Figure 2.14, Chapter 2). UKCTOCS DCS variant calls, from 20 pre-AML and 20 controls are shown **a.** before *in silico* noise correction and post-processing. **b.** after *in silico* noise correction and post-processing.

4.5.5 Testing the position-specific method on Myeloid DNA Reference Standards.

Whilst the *in silico* noise correction method appears to perform well on the UKCTOCS samples, testing it on the Myeloid DNA Reference Standard (Horizon Discovery Ltd.) samples revealed poor performance, even in the undiluted samples where variants were at $\geq 5\%$ VAF. Further analysis revealed that this is because the method struggles to identify real variants when two or more samples at a position have a real variant at approximately the same VAF. This was the case for the Myeloid DNA Reference Standards, which were all sequenced in replicate on the same NovaSeq 6000 S4 flow cell. The variants in the Myeloid Reference Standards were all COSMIC hotspot sites and so an iterative beta-binomial approach was used, with exclusion of the highest VAF sample at each iteration. However, even with exclusion of the highest VAF sample (e.g. 100% Myeloid Reference Standard, replicate 1), the replicate was still present (e.g. 100% Myeloid Reference Standard, replicate 2) and so when the beta-binomial distribution was fitted, the dispersion was over-estimated, resulting in all of the samples being called as errors (Figure 4.21). Whilst this situation might be expected to be rare with the real data, we tried to minimise the risk of this by sequencing timepoints from the same person on different flow-cells. Future improvements to the *in silico* noise correction method will involve using a prior distribution for δ (which penalises large δ), particularly at COSMIC hotspot sites, which will also help to avoid this issue.

4.6 Detecting indels and FLT3-ITD mutations

Developing an error model for indels is more complicated than for single nucleotide variants, given the variability in length and base insertions/ deletions at a particular site. We plan to explore this further, but in the meantime, we chose to be quite conservative with which indels we called as real. Indels were excluded if they were detected in <4 DCS reads, unless they were a known NPM1 exon 12

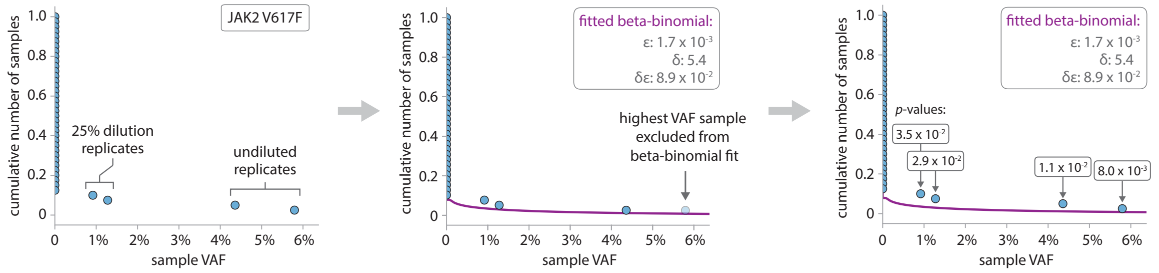


Figure 4.21 Difficulty with calling variants when real variants with similar high VAF are present. All Myeloid Reference Standard dilution replicates were sequenced on the same NovaSeq 6000 S4 flow-cell. When the position-specific error model was applied to all samples on the same flow-cell, an iterative beta-binomial approach was used for the Myeloid Reference Standard real variant sites, because they were all COSMIC hotspots (e.g. JAK2 V617F shown). The highest VAF sample was excluded and a beta-binomial distribution (purple line) fitted to the remaining samples. Because a replicate of the highest VAF sample was present, however, an over-dispersed beta-binomial distribution was fitted. When the *p*-values were calculated for all the samples, none of them were $< 6 \times 10^{-6}$ and so none of them were called as real.

frameshift hotspot mutation. Indels were also excluded if they were seen in ≥10% of people, unless the variant had been observed in haematopoietic and lymphoid tissues in COSMIC v92[166]. Using information from multiple timepoints, indels were excluded if they were not detected at >4% VAF at ≥1 timepoint and indels that were detected at some timepoints, but not others, were filtered using the same method as for SNVs (eqs. 4.6-4.8).

FLT3-ITD mutations were called using Pindel[184] and were detected in 2-3 SSCS reads of 5 of our final timepoint UKCTOCS pre-AML samples. Two of these cases also had a FLT3-ITD mutation detected in their penultimate timepoint sample, but on further inspection, these tandem duplications were at different positions and of different lengths to those detected in the final timepoint sample. We were therefore concerned that the majority of these FLT3-ITD calls were errors and so, pending confirmatory RT-PCR, we only called a FLT3-ITD if it was detected in both the SSCS and DCS. This meant we only called a FLT3-ITD mutation in one pre-AML case.

4.7 Detecting mosaic chromosomal alterations (mCAs)

Our targeted panel for mCAs, KMT2A-PTD and chromosomal rearrangements contains a ‘SNP backbone’ which targets ~10,000 regions (each 120bp) spaced ~280kb across the genome and each containing ≥1 common SNPs (at MAF 0.4-0.45 in 1000 genomes[168]). A custom Python script was written to enable detection of mCAs from the SSCS reads using the SNP ‘B-allele frequencies’ (BAF) and read depths ($\log_2$R ratios) in these targeted regions.

Calculating $\log_2$R ratios (LRR) and B-allele frequencies (BAF)

For each sample, the average read depth across each 120 bp targeted region (SNP region) was calculated. The average read depth of each SNP region was then normalised to the average read depth across all targeted regions in the sample, to create ‘sample normalised read depths’ for each SNP region. To account for inter-region variation in read depth coverage, due to variability in capture efficiency, the ∼20 control samples from the same sequencing lane were used as a ‘panel of normals’ to calculate the average normalised read depth for each SNP region. The mean coefficient of variation (CV) in read depths was calculated across the ‘panel of normals’, for each chromosome, and controls were excluded from the ‘panel of normals’ if their read depth CV for ≥ 2 chromosomes was more than 1.5 standard deviations above the mean CV. This helped to prevent any samples that had a significant copy number alteration (i.e. gain or loss event) being included in the ‘panel of normals’. Log_2R ratios (LRR) were then calculated for each SNP region by comparing the region’s ‘sample normalised read depth’ to the average normalised read depth for that region from the ‘panel of normals’.

The VAF of all targeted SNPs, as well as any SNPs >1% MAF in 1000 genomes[168] that were also covered by the panel, were calculated (from read depth/total depth) and were plotted across the chromosomes to visualise the B-allele frequency (BAF).

Identifying mCAs as either gain, loss or CN-LOH events

Once BAFs and $\log_2$R ratios (LRR) had been calculated, it was possible to identify regions in which either copy number alterations (i.e loss or gain events) or copy-neutral loss of heterozygosity (CN-LOH) events had occurred. Loss and gain events result in deviations in both LRR and BAF whereas CN-LOH events result in BAF deviations without a change in LRR (because there is no change in the amount of genetic material) (Figure 4.22).

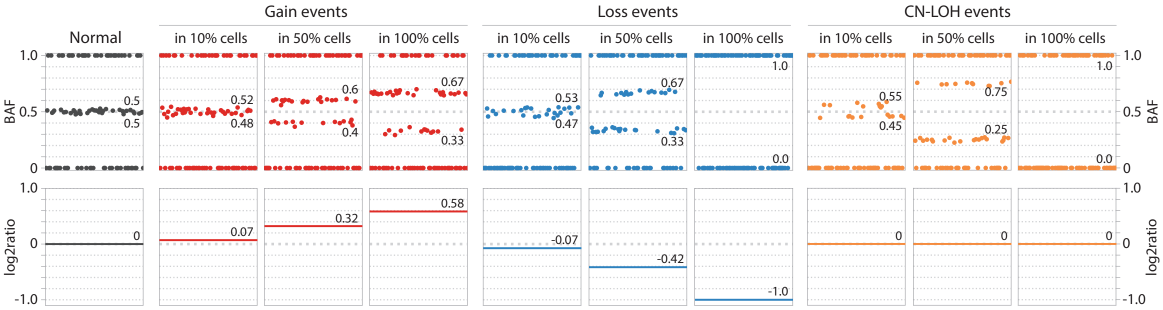


Figure 4.22 LRR and BAF deviations for mCA detection. Normal regions (black plot) contain SNPs with B-allele frequencies (BAFs) of 0 (homozygous AA alleles), 0.5 (heterozygous AB alleles) and 1.0 (homozygous BB alleles) and no change in read depth ($\log_2$ratio of 0). Regions with gain events (red plots) show deviations in BAF, up to a maximum of ±0.17 if a duplication event is in 100% of cells. Gain events result in an increase in read depth ($\log_2\text{ratio} > 0$), due to the extra genetic material. Regions with loss events (blue plots) or CN-LOH events (orange plots) also show deviations in BAF and can both have full loss of heterozygosity if the loss or CN-LOH affects 100% of cells. Loss and CN-LOH events can be distinguished by the read depths, which are decreased in loss events ($\log_2\text{ratio} < 0$) due to loss of genetic material, but are unchanged in CN-LOH events ($\log_2\text{ratio} = 0$). Data generated for schematic using simulated samples (100 SNPs per region with mean read depth across ‘panel’ of 1000 reads).

Determining length and cell fraction of mCAs

Because BAF deviations could occur due to inadvertent somatic variants at common SNP sites or variable bait capture, we required BAF deviations (or lack of heterozygosity) in $\geq$5 consecutive SNPs for an mCA to be called. With SNPs spaced $\sim$ 280kb apart, this meant the smallest mCA we could expect to detect would be $\sim$ 1.5 MB. The 'start' of the mCA was taken as the coordinate of the first BAF-deviated SNP and the 'end' coordinate was taken as the coordinate of the last BAF-deviated SNP in the affected region.

The proportion of cells ('cell fraction') harbouring the mCA was calculated from the heterozygous BAFs as detailed in Table 4.6. If an mCA was detected at 100% cell fraction in both the latest and earliest timepoint sample from an individual, then it was deemed likely to be germline.

Table 4.6 Expected LRR, heterozygous BAFs and cell fractions for autosomal somatic mCAs. mCA cell fraction (p) can be calculated directly from the heterozygous BAFs (μ_1 and μ_2). Adapted from Jacobs et al.[30]

Somatic mCA (autosomal)	**LRR**	**Heterozgous BAFs (μ_1, μ_2)**	**mCA cell fraction (p)**
Gain	$\log_2\left(\frac{2+p}{2}\right)$	$0.5 \pm \frac{p}{2(2+p)}$	$\frac{2(\mu_2-\mu_1)}{1-(\mu_2-\mu_1)}$
Loss	$\log_2\left(\frac{2-p}{2}\right)$	$0.5 \pm \frac{p}{2(2-p)}$	$\frac{2(\mu_2-\mu_1)}{1+\mu_2-\mu_1}$
CN-LOH	0	$0.5 \pm \frac{p}{2}$	$\mu_2 - \mu_1$

Using phasing information to call mCAs at low cell fractions

Because mCA calling relies on the ability to detect deviations in heterozygous BAFs, the noisier the BAF measurements (e.g. due to low sequencing read depth), the harder it will be to detect subtle BAF deviations associated with low cell fraction mCAs (e.g. gain event in 10% cells in Figure 4.22). Haplotype phasing, which involves identifying SNPs which lie on the same chromosome and therefore have the same direction of BAF deviation, can be used to improve the sensitivity of low cell fraction mCA calling. This approach has been used to call mCAs at cell fractions as low as 1% in UK Biobank participants[34,35] by utilising long-range phase information generated from long identical-by-descent (IBD) tracts shared among distantly related individuals[189]. This information is unfortunately not available for the UKCTOCS participants. However, one of the benefits afforded by having longitudinal samples is the ability to use large BAF deviations from higher-cell fraction mCAs, detected at timepoints closer to AML diagnosis, to identify which SNPs lie on the same chromosome in the affected region of interest. This phasing information can then be applied to the same individual's samples from earlier timepoints, allowing a much higher sensitivity for detection of the mCA when it is at lower cell fraction (Figure 4.23). Using this approach we were able to call mCAs at cell fractions as low as 0.2%.

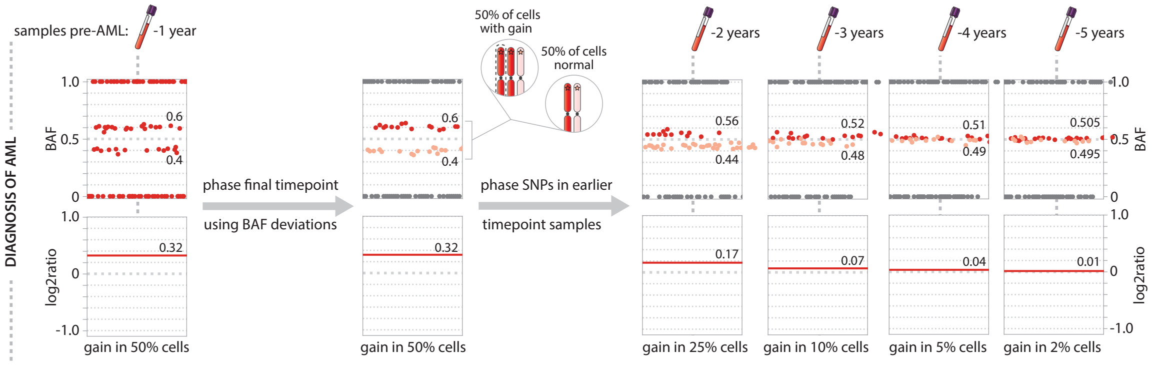


Figure 4.23 Phasing SNPs for detection of low cell fraction mCAs. mCAs detected at high enough cell fraction to allow differentiation of the heterozygous SNPs that have deviated above and below 0.5 BAF enable SNPs to be 'phased' (e.g. in the final timepoint 'sample' pre-AML diagnosis shown on the left). This phasing information can then be applied to earlier timepoint samples (shown on the right) and provides higher sensitivity for detection of smaller BAF deviations and therefore higher sensitivity for detection of mCAs at lower cell fraction. Data generated for schematic using a simulated sample (100 SNPs per region with mean read depth across 'panel' of 1000 reads).

4.8 Detecting KMT2A partial tandem duplications (KMT2A-PTD)

Partial tandem duplications in KMT2A (KMT2A-PTD) most commonly involve exons 2 or 3 and span through exon 9 to 11[171]. Exon 27, which is the largest exon in KMT2A, is characteristically not involved. Therefore to detect KMT2A-PTD events a custom Python script was written (using pysam[190] pileup) to calculate the mean read depth across each targeted KMT2A exon. These mean depths were then normalised to the mean read depth across KMT2A exon 27 and the exon 3:27 ratio (R) calculated. The fraction of cells harbouring the KMT2A-PTD can then be calculated as:

$$\text{cell fraction} = 2(R-1) \tag{4.9}$$

A KMT2A-PTD affecting one KMT2A allele in 100% of cells should be easy to detect, with a mean read depth ratio (R) of 1.5 (Figure 4.24a). At lower cell fractions, however, the ratio becomes quite small and may become harder to detect (Figure 4.24b), particularly if the read depths within each exon are highly variable. This is a limitation of our KMT2A-PTD detection method and means we may not detect KMT2A-PTD events pre-AML diagnosis if they are at low cell fraction. KMT2A-PTD are effectively small gain events and so SNP BAFs could be used to improve sensitivity, akin to their use in mCA detection. There are, however, only 10 SNPs at >1% MAF in 1000 genomes[168] across our targeted KMT2A regions and none of these are in exon 3.

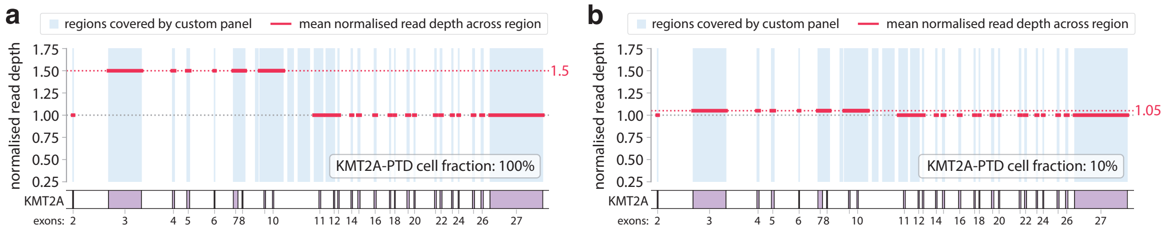


Figure 4.24 Schematic showing the effect of KMT2A-PTD on exon 3: exon 27 read depth ratios. **a.** KMT2A-PTD involving exons 3-10 in 100% of cells results in an exon 3: exon 27 read depth ratio of 1.5. **b.** KMT2A-PTD involving exons 3-10 in 10% of cells results in an exon 3: exon 27 read depth ratio of 1.05.

4.9 Developing a caller for chromosomal rearrangments

A number of software packages are available for the detection of chromosomal rearrangements from short-read NGS data, e.g BreakDancer[191], CREST[192], Delly (EMBL)[193], Lumpy[194], Manta (Illumina)[195] and GRIDDS[196]. These packages typically use one of three different methods to detect chromosomal rearrangements: i) discordant read pairs, ii) split reads or iii) combination methods with or without local re-assembly.

The ‘discordant read pair’ method involves the identification of read pairs that are mapped in an abnormal orientation or whose mapped interval is significantly different to the expected library insert size (Figure 4.25a). A key limitation of this method is its poor breakpoint resolution as well as low sensitivity for low VAF events. BreakDancer[191] is an example of a caller that uses this method.

The ‘split reads’ method looks for regions where there are increased numbers of ‘soft-clipped’ incompletely mapped reads, which may be suggestive of a chimeric read spanning a rearrangement breakpoint (Figure 4.25b). The soft-clipped reads can then be assembled in to a contig spanning the translocation breakpoint, which can then be aligned back to the reference genome (‘local re-assembly’) to determine the breakpoint positions. The benefit of the ‘split reads’ method is its ability to infer breakpoint positions at single-nucleotide resolution and it has higher sensitivity than the ‘discordant reads’ method, although high sequencing depth is typically needed to obtain sufficient reads overlapping the breakpoint. CREST[192] is an example of a caller that uses this method.

Combination methods typically involve the ‘discordant read pair’ method first, to predict the presence of chromosomal rearrangements, and then the ‘split read’ method next, with or without local reassembly, to refine the breakpoint positions. The advantage of local reassembly is its ability to reconstruct novel rearrangements, although low VAF events can be difficult to detect as there may not be sufficient variant reads present for contig assembly[197]. Callers that utilise a combination of methods as well as local realignment, such as Manta[195] and GRIDDS[196] generally have higher sensitivity and lower false discovery rates[197] than callers that use a single method, e.g. BreakDancer[191], which has a sensitivity of only 20-30% even when the rearrangement is at 100% VAF[197]. Existing packages generally struggle with low VAF chromosomal rearrangements, with sensitivity dropping to ∼60% at 10%

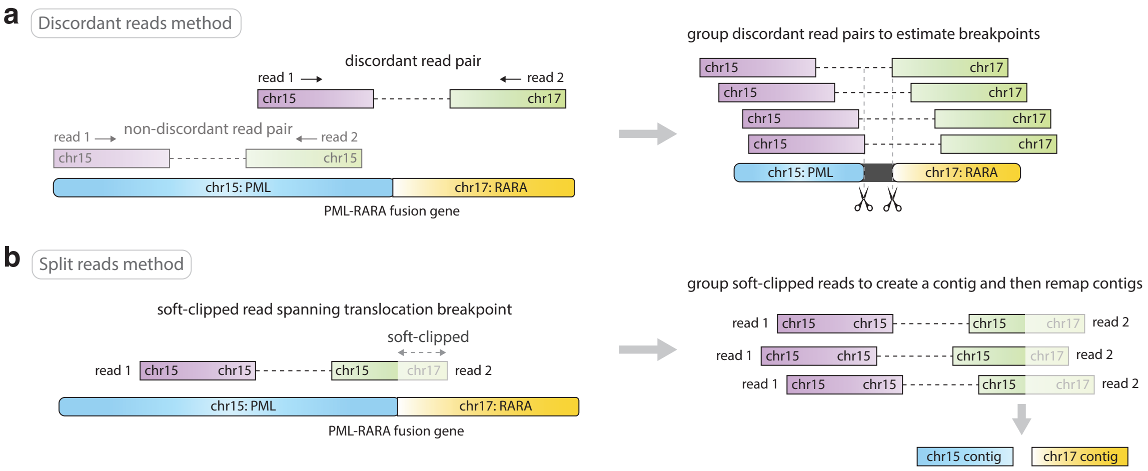


Figure 4.25 Common methods utilised for detection of chromosomal rearrangements. **a.** 'Discordant reads' method identifies read pairs that are mapped in an abnormal orientation or whose mapped interval is significantly different to the expected library insert size. **b.** 'Split reads' method looks for regions where there are increased number of 'soft-clipped' incompletely mapped reads. Many translocation caller packages use a combination of these methods.

VAF in Manta[195], Lumpy[194] and GRIDDs[196], although this is better than Delly[193], which reportedly has zero sensitivity at VAF <10%. Detection of intra-chromosomal translocations is also reportedly problematic, with BreakDancer[191], Delly[193], Lumpy[194], Manta[195] and GRIDDs[196] all missing ∼50% of intra-chromosomal translocation events in a head-to-head performance assessment[197].

Given the limitations of these software packages, and our desire for a caller that could reliably detect chromosomal rearrangements at low VAF (e.g. in the years pre-AML diagnosis) as well as intra-chromosomal translocations, we sought to develop our own custom caller.

4.9.1 Types of chromosomal rearrangement to detect

We first sought to understand the different types of chromosomal rearrangement we needed to detect. Chromosomal rearrangements can occur between non-homologous chromosomes (e.g. t(8;21)), between homologous chromosomes (e.g. t(16;16)) or within a chromosome (e.g. inv(16)). Depending on whether the rearrangement is between two p-arms (or q-arms) or between a p- and a q-arm, the relocated region can end up in its original orientation or it can become inverted. For example in t(8;21) part of RUNX1, which is on chromosome 21q, is relocated without inversion to join next to part of the RUNX1T1 gene, which is on chromosome 8q. In contrast, in t(9;11) part of KMT2A, which is on chromosome 11q, is relocated and inverted to join next to the MLLT3 gene, which is on chromosome 9p. Chromosomal rearrangements can therefore be categorised into four different classes (Figure 4.26): i) rearrangement between non-homologous chromosomes with the relocated regions remaining in their original orientation, ii) rearrangement between homologous chromosomes (or within a chromosome) with the relocated regions remaining in their original orientation, iii) rearrangement between non-homologous chromosomes with inversion of the relocated regions, and

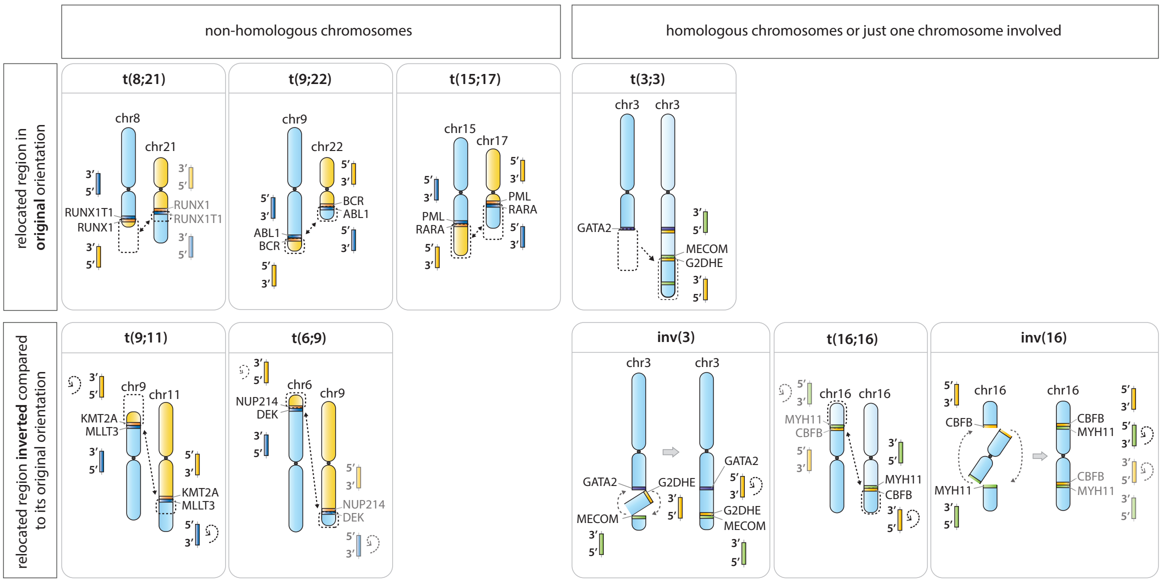

non-homologous chromosomes
homologous chromosomes or just one chromosome involved
relocated region in original orientation
relocated region inverted compared to its original orientation
t(8;21)
chr8
chr21
RUNX1T1
RUNX1
t(9;22)
chr9
chr22
BCR
ABL1
t(15;17)
chr15
chr17
PML
RARA
t(3;3)
chr3
GATA2
MECOM
G2DHE
t(9;11)
KMT2A
MLLT3
chr11
t(6;9)
chr6
NUP214
DEK
inv(3)
t(16;16)
chr16
MYH11
CBFB
inv(16)
3′
5′

Figure 4.26 Categories of chromosomal rearrangement. Chromosomal rearrangements can be grouped into one of four categories, depending on whether the rearrangement involves non-homologous or homologous chromosomes and whether or not the relocated regions become inverted compared to their original orientation. The chromosomal rearrangements targeted by our custom panel include rearrangements in each of these four categories as shown.

iv) rearrangement between homologous chromosomes (or within a chromosome) with inversion of the relocated regions. The AML-associated chromosomal rearrangements targeted by our custom panel include rearrangements in each of these categories (Figure 4.26) and so it was important that our caller could reliably detect all four categories.

4.9.2 Characteristics of paired end reads spanning chromosomal rearrangements

We next sought to understand the features we would expect to see in paired end sequencing data for each of these four categories of chromosomal rearrangement. A custom Python script was written to generate simulated ‘samples’, each containing a different chromosomal rearrangement from each of the four categories, at a defined VAF (Appendix C.2.1). We used the mapped SSCS BAM files from these samples, which were aligned using BWA[180], to understand the features that characterise reads that span breakpoint regions to ensure that our caller could capture all of these reads. Read features in the BAM file were systematically explored using pysam[190] and Integrative Genomics Viewer (IGV)[198].

Primary alignments and supplementary alignments

If BWA[180] encounters a read that maps to two different chromosomes (or different regions of the same chromosome), e.g. chr22 at one end and chr9 at the other, it will soft-clip the shorter aligned sequence and report the longer aligned sequence as the ‘primary alignment’ for that read (Figure 4.27). If the shorter aligned sequence is longer than 30 bp, however, BWA[180] also outputs an alternatively mapped read in which the longer aligned sequence is soft-clipped and the shorter aligned sequence is reported as the ‘supplementary alignment’. This means that as long as at least one of the reads in the pair overlaps the rearrangement breakpoint by >30bp, there will always be a primary alignment and a supplementary alignment present for the pair of reads. The minimum sequence length of 30bp can be adjusted, although setting this too short will increase the chance of mis-mapping information in the supplementary alignment. Supplementary reads can be identified in the BAM file by their SAM FLAG which is usually four digits starting with a ‘2’ (e.g. 2145 and 2193 in Figure 4.27)[199].

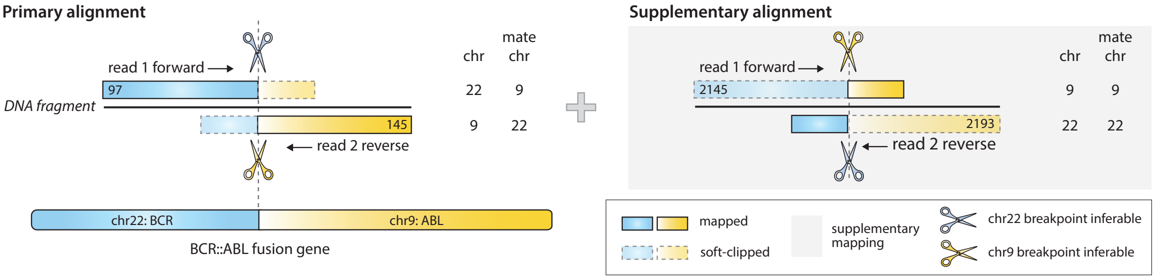


Figure 4.27 Primary and supplementary alignments of a pair of reads spanning a t(9;22) BCR::ABL breakpoint region. If a read maps to two different chromosomes, or chromosomal regions, BWA[180] assigns the mapping of the longer aligned sequence as the read’s ‘primary alignment’ and ‘soft-clips’ the shorter sequence that maps to another region. It also outputs a ‘supplementary alignment’ of the mapping of the shorter aligned sequence with ‘soft-clipping’ of the longer sequence that maps to another region. The alignment information of a read’s ‘mate’ is that of the ‘mate’ in the primary alignment (e.g. the ‘mate’ of the supplementary read 1 forward shown is the primary read 2 reverse read). The number shown within the read is the read’s SAM flag.

Discordant and concordant reads

Read pairs can be defined as ‘discordant’ or ‘concordant’. ‘Discordant reads’ are those whose read pair ‘mate’ maps to a different chromosome or chromosomal region (e.g. read 1 is mapped to chr22 and read 2 is mapped to chr9 in the primary alignment reads in Figure 4.27). ‘Concordant reads’ are those whose read pair ‘mate’ maps to the same chromosomal region (e.g. both read 1 and read 2 map to chr9). It is important to be aware that, for supplementary alignments, the ‘mate’ referred to in the BAM alignment information is the primary alignment mate, not the supplementary alignment mate (i.e. the ‘mate’ of a read 1 supplementary alignment is the read 2 primary alignment). For example, in Figure 4.27, although the supplementary read pairs are mapping to different chromosomes, they are mapping to the same chromosome as their primary alignment mate and so these supplementary reads would be classified as ‘concordant’.

The method for identifying discordant read pairs differs slightly depending on the category of chromosomal rearrangement. For rearrangements involving non-homologous chromosomes, discordant reads can simply be identified as those which map to different chromosomes. For rearrangements involving homologous chromosome(s), discordant read pairs map to the same chromosome, but the ‘template length’ (reported in the BAM alignment information) is significantly larger than the expected library insert size. For non-homologous or homologous rearrangements that involve inversion of the relocated segment, the discordant reads have the additional feature of being mapped in the same direction (e.g. read 1 forward and read 2 forward).

Identifying breakpoint partners

Depending on the length of the DNA fragment and where the rearrangement breakpoint lies along the length of the fragment, several different read pair scenarios can arise (Figures 4.28, 4.29 and 4.30). The scenario to which a read pair belongs can be determined by the reads' SAM FLAGS and BAM alignment information as shown in Figure 4.28 (chromosomal rearrangements without inversion), Figure 4.29 (chromosomal rearrangements with inverted segment on the 'left') and Figure 4.30 (chromosomal rearrangements with inverted segment on the 'right').

In scenario 1, where the DNA fragment is approximately the same length as the read, the pair of reads both have their predominant mapping on the same chromosomal region. The primary alignments will therefore be 'concordant' and the supplementary alignments will be 'discordant' ('scenario 1 read pairs' in Figures 4.28, 4.29 and 4.30). In this scenario, one of the breakpoint partners can be inferred from the mapped location of the primary alignment and the other breakpoint partner can be inferred from the mapped location of the supplementary alignment.

In scenario 2, where the DNA fragment is longer than the read length but the read pairs still overlap, the pair of reads have their predominant mapping on different chromosomes. The primary alignments will therefore be 'discordant' and the secondary alignments will be 'concordant' ('scenario 2 read pairs' in Figures 4.28, 4.29 and 4.30). In this scenario, if both reads overlap the breakpoint region then both breakpoint partners can be inferred from either the primary or supplementary alignment (Figure 4.28, read pair d, Figures 4.29 and 4.30, read pair e). If only one of the reads overlaps the breakpoint then one breakpoint partner can be inferred from the primary alignment and the other breakpoint partner from the supplementary alignment (Figure 4.28, read pair e, Figures 4.29 and 4.30, read pairs f-g).

In scenario 3 only one, or none, of the breakpoint partners can be inferred from the reads, either because the DNA fragment is longer than both reads, or because the breakpoint region lies towards the end of the DNA fragment ('scenario 3 read pairs', Figures 4.28, 4.29 and 4.30). When neither of the reads span the breakpoint (e.g. because the DNA fragment is very long) it is only possible to infer bounds on the breakpoint regions (Figure 4.28, read pairs g, Figures 4.29 and 4.30, read pair j). When only one of the breakpoints can be inferred (e.g. because the soft-clipped region is too short to have formed a supplementary alignment) the only way to infer the other breakpoint region would be to remap the short soft-clipped region (Figure 4.28, read pairs f, h, i, Figures 4.29 and 4.30, read pairs h, i, k, l).

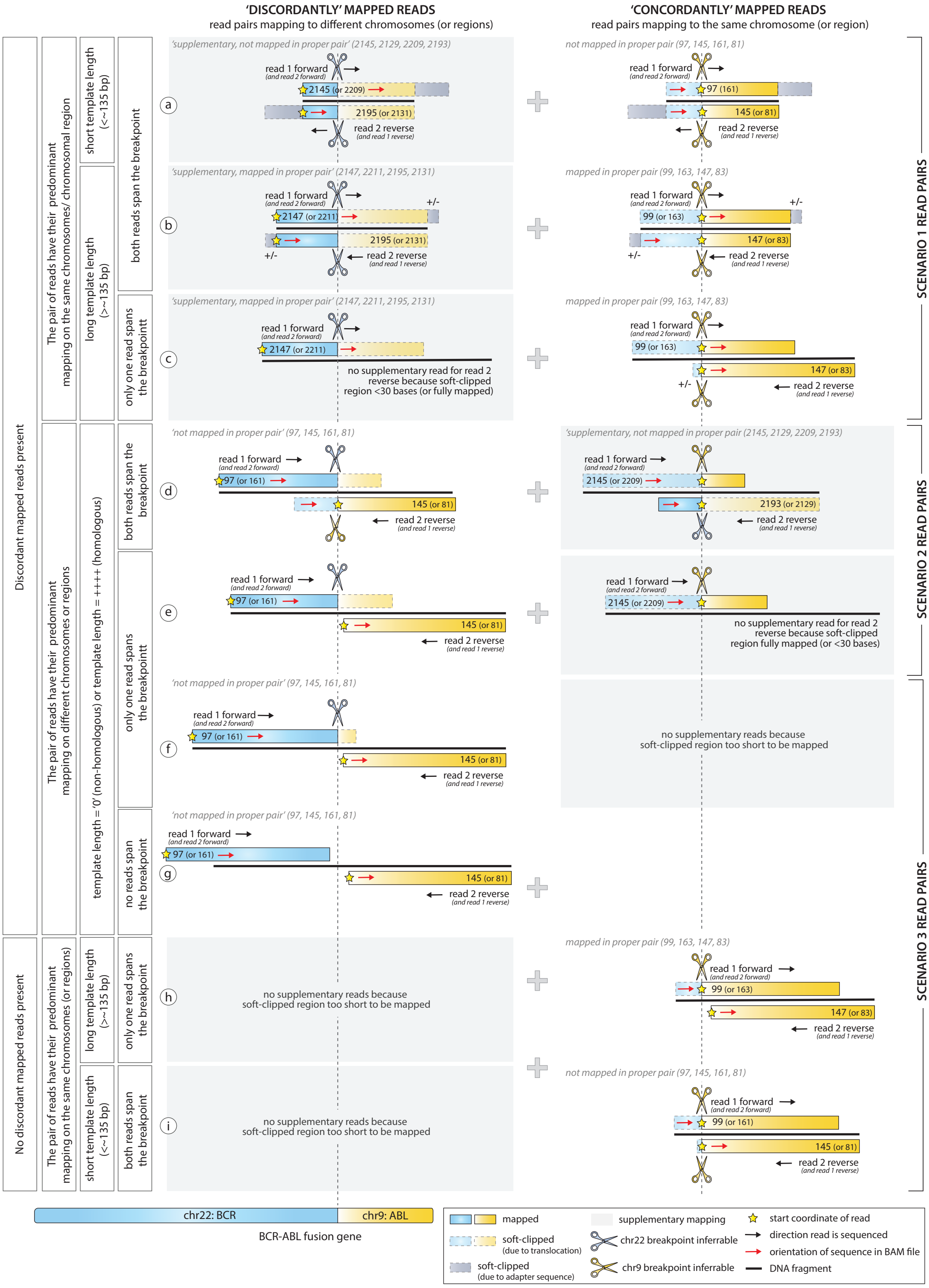
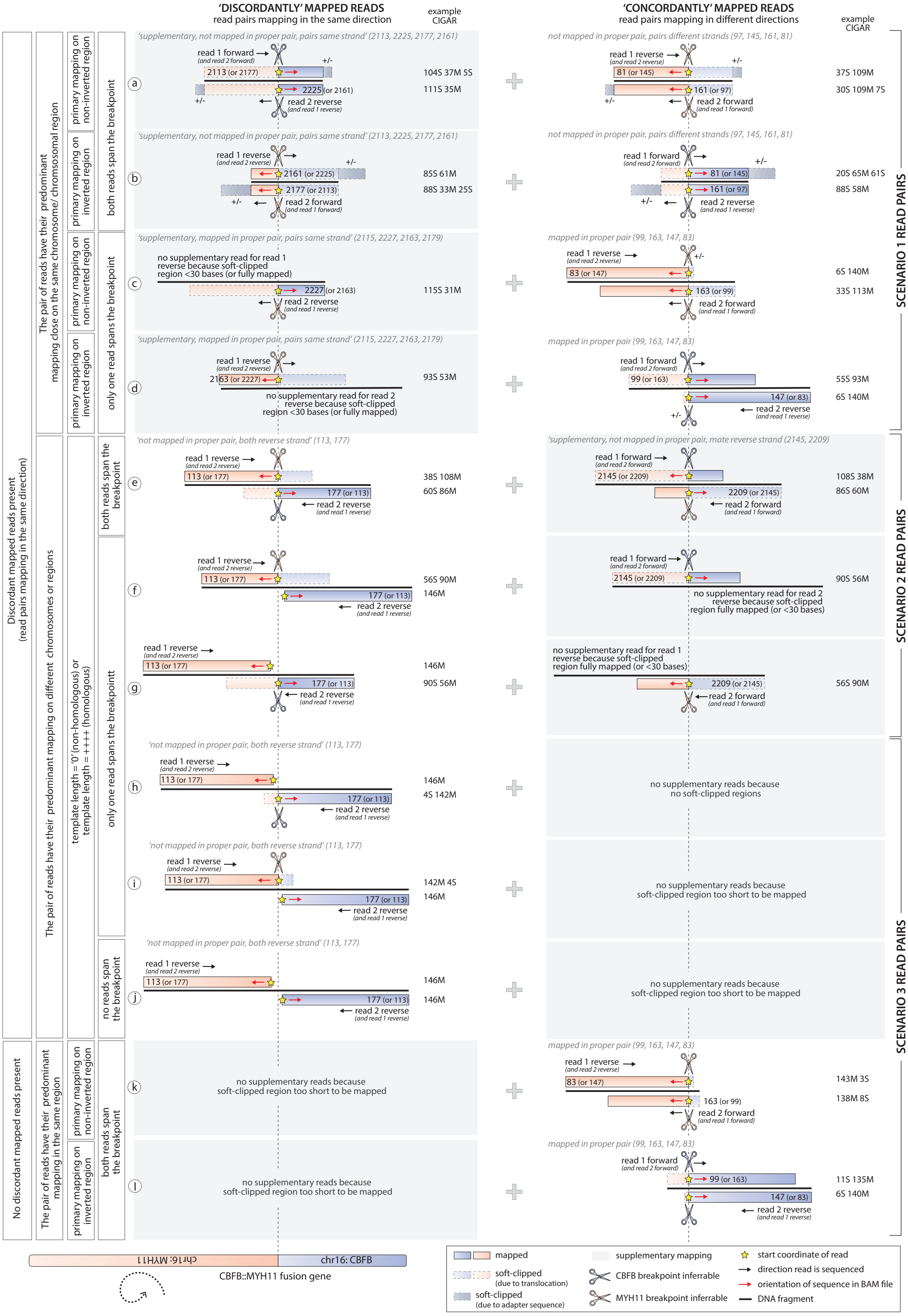

'DISCORDANTLY' MAPPED READS
read pairs mapping to different chromosomes (or regions)
'CONCORDANTLY' MAPPED READS
read pairs mapping to the same chromosome (or region)
Discordant mapped reads present
No discordant mapped reads present
The pair of reads have their predominant mapping on the same chromosomes/ chromosomal region
The pair of reads have their predominant mapping on different chromosomes or regions
The pair of reads have their predominant mapping on the same chromosomes (or regions)
short template length (<~135 bp)
long template length (>~135 bp)
template length = '0' (non-homologous) or template length = ++++ (homologous)
both reads span the breakpoint
only one read spans the breakpoint
both reads span the breakpoint
only one read spans the breakpointt
no reads span the breakpoint
'supplementary, not mapped in proper pair' (2145, 2129, 2209, 2193)
'supplementary, mapped in proper pair' (2147, 2211, 2195, 2131)
'not mapped in proper pair' (97, 145, 161, 81)
not mapped in proper pair (97, 145, 161, 81)
mapped in proper pair (99, 163, 147, 83)
read 1 forward (and read 2 forward)
read 2 reverse (and read 1 reverse)
2145 (or 2209)
2195 (or 2131)
2147 (or 2211)
2193 (or 2129)
97 (or 161)
145 (or 81)
99 (or 163)
147 (or 83)
+/-
a
b
c
d
e
f
g
h
i
no supplementary read for read 2 reverse because soft-clipped region <30 bases (or fully mapped)
no supplementary read for read 2 reverse because soft-clipped region fully mapped (or <30 bases)
no supplementary reads because soft-clipped region too short to be mapped
SCENARIO 1 READ PAIRS
SCENARIO 2 READ PAIRS
SCENARIO 3 READ PAIRS
chr22: BCR
chr9: ABL
BCR-ABL fusion gene
mapped
supplementary mapping
start coordinate of read
soft-clipped (due to translocation)
chr22 breakpoint inferrable
direction read is sequenced
orientation of sequence in BAM file
soft-clipped (due to adapter sequence)
chr9 breakpoint inferrable
DNA fragment

Figure 4.28 Characteristics of read pairs that span chromosomal rearrangement breakpoint regions, for rearrangements that do not involve inversion of the relocated segments (either non-homologous or homologous). In scenario 1 read pairs, one of the breakpoint partners can be inferred from the primary alignment and the other breakpoint partner can be inferred from the supplementary alignment. In scenario 2 read pairs, one or both breakpoint partners can be inferred from either the primary alignment or the supplementary alignment. In scenario 3 read pairs, only one, or none of the breakpoint partners can be inferred from the reads.



'DISCORDANTLY' MAPPED READS
read pairs mapping in the same direction
example CIGAR
'CONCORDANTLY' MAPPED READS
read pairs mapping in different directions
example CIGAR
Discordant mapped reads present
(read pairs mapping in the same direction)
No discordant mapped reads present
The pair of reads have their predominant mapping close on the same chromosome/ chromosomal region
The pair of reads have their predominant mapping on different chromosomes or regions
The pair of reads have their predominant mapping in the same region
primary mapping on non-inverted region
primary mapping on inverted region
template length = '0' (non-homologous) or template length = ++++ (homologous)
both reads span the breakpoint
only one read spans the breakpoint
no reads span the breakpoint
a
b
c
d
e
f
g
h
i
j
k
l
'supplementary, not mapped in proper pair, pairs same strand' (2113, 2225, 2177, 2161)
'supplementary, mapped in proper pair, pairs same strand' (2115, 2227, 2163, 2179)
'not mapped in proper pair, both reverse strand' (113, 177)
not mapped in proper pair, pairs different strands (97, 145, 161, 81)
mapped in proper pair (99, 163, 147, 83)
'supplementary, not mapped in proper pair, mate reverse strand (2145, 2209)
read 1 forward
(and read 2 forward)
read 1 reverse
(and read 2 reverse)
read 2 forward
(and read 1 forward)
read 2 reverse
(and read 1 reverse)
+/-
2113 (or 2177)
2225 (or 2161)
2161 (or 2225)
2177 (or 2113)
2227 (or 2163)
2163 (or 2227)
113 (or 177)
177 (or 113)
81 (or 145)
161 (or 97)
83 (or 147)
163 (or 99)
99 (or 163)
147 (or 83)
2145 (or 2209)
2209 (or 2145)
104S 37M 5S
111S 35M
85S 61M
88S 33M 25S
115S 31M
93S 53M
38S 108M
60S 86M
56S 90M
146M
90S 56M
4S 142M
142M 4S
37S 109M
30S 109M 7S
20S 65M 61S
88S 58M
6S 140M
33S 113M
55S 93M
108S 38M
86S 60M
143M 3S
138M 8S
11S 135M
no supplementary read for read 1 reverse because soft-clipped region <30 bases (or fully mapped)
no supplementary read for read 2 reverse because soft-clipped region <30 bases (or fully mapped)
no supplementary read for read 2 reverse because soft-clipped region fully mapped (or <30 bases)
no supplementary read for read 1 reverse because soft-clipped region fully mapped (or <30 bases)
no supplementary reads because no soft-clipped regions
no supplementary reads because soft-clipped region too short to be mapped
SCENARIO 1 READ PAIRS
SCENARIO 2 READ PAIRS
SCENARIO 3 READ PAIRS
chr16: MYH11
chr16: CBFB
CBFB::MYH11 fusion gene
mapped
soft-clipped (due to translocation)
soft-clipped (due to adapter sequence)
supplementary mapping
CBFB breakpoint inferrable
MYH11 breakpoint inferrable
start coordinate of read
direction read is sequenced
orientation of sequence in BAM file
DNA fragment

Figure 4.29 Characteristics of read pairs that span chromosomal rearrangement breakpoint regions, for rearrangements that involve inversion of the relocated segment on the 'LEFT' of the breakpoint fusion (either non-homologous or homologous). In scenario 1 read pairs, one of the breakpoint partners can be inferred from the primary alignment and the other breakpoint partner can be inferred from the supplementary alignment. In scenario 2 read pairs, one or both breakpoint partners can be inferred from either the primary alignment or the supplementary alignment. In scenario 3 read pairs only one, or none of the breakpoint partners can be inferred from the reads.

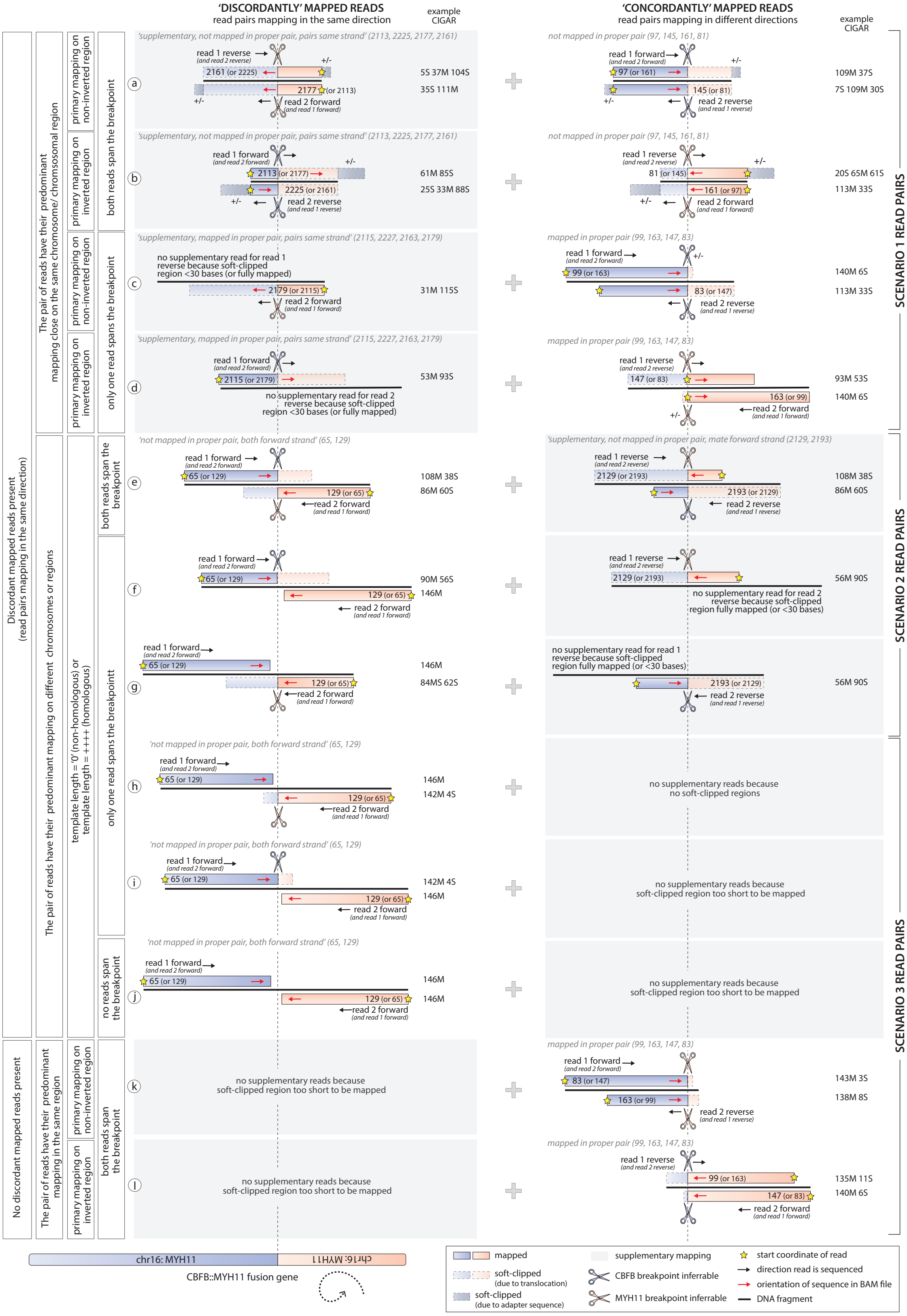

'DISCORDANTLY' MAPPED READS
read pairs mapping in the same direction
example CIGAR
'CONCORDANTLY' MAPPED READS
read pairs mapping in different directions
example CIGAR
Discordant mapped reads present (read pairs mapping in the same direction)
The pair of reads have their predominant mapping close on the same chromosome/ chromosomal region
The pair of reads have their predominant mapping on different chromosomes or regions
template length = '0' (non-homologous) or template length = ++++ (homologous)
No discordant mapped reads present
The pair of reads have their predominant mapping in the same region
primary mapping on non-inverted region
primary mapping on inverted region
both reads span the breakpoint
only one read spans the breakpoint
no reads span the breakpoint
'supplementary, not mapped in proper pair, pairs same strand' (2113, 2225, 2177, 2161)
'supplementary, mapped in proper pair, pairs same strand' (2115, 2227, 2163, 2179)
'not mapped in proper pair, both forward strand' (65, 129)
not mapped in proper pair (97, 145, 161, 81)
mapped in proper pair (99, 163, 147, 83)
'supplementary, not mapped in proper pair, mate forward strand (2129, 2193)
a
5S 37M 104S
35S 111M
109M 37S
7S 109M 30S
b
61M 85S
25S 33M 88S
20S 65M 61S
113M 33S
c
no supplementary read for read 1 reverse because soft-clipped region <30 bases (or fully mapped)
31M 115S
140M 6S
113M 33S
d
no supplementary read for read 2 reverse because soft-clipped region <30 bases (or fully mapped)
53M 93S
93M 53S
140M 6S
e
108M 38S
86M 60S
108M 38S
86M 60S
f
90M 56S
146M
56M 90S
no supplementary read for read 2 reverse because soft-clipped region fully mapped (or <30 bases)
g
146M
84MS 62S
no supplementary read for read 1 reverse because soft-clipped region fully mapped (or <30 bases)
56M 90S
h
146M
142M 4S
no supplementary reads because no soft-clipped regions
i
142M 4S
146M
no supplementary reads because soft-clipped region too short to be mapped
j
146M
146M
no supplementary reads because soft-clipped region too short to be mapped
k
no supplementary reads because soft-clipped region too short to be mapped
143M 3S
138M 8S
l
no supplementary reads because soft-clipped region too short to be mapped
135M 11S
140M 6S
read 1 forward
read 1 reverse
read 2 forward
read 2 reverse
(and read 1 forward)
(and read 1 reverse)
(and read 2 forward)
(and read 2 reverse)
+/-
SCENARIO 1 READ PAIRS
SCENARIO 2 READ PAIRS
SCENARIO 3 READ PAIRS
chr16: MYH11
chr16: MYH11
CBFB::MYH11 fusion gene
mapped
soft-clipped (due to translocation)
soft-clipped (due to adapter sequence)
supplementary mapping
CBFB breakpoint inferrable
MYH11 breakpoint inferrable
start coordinate of read
direction read is sequenced
orientation of sequence in BAM file
DNA fragment

Figure 4.30 Characteristics of read pairs that span chromosomal rearrangement breakpoint regions, for rearrangements that involve inversion of the relocated segment on the 'RIGHT' of the breakpoint fusion (either non-homologous or homologous). In scenario 1 read pairs, one of the breakpoint partners can be inferred from the primary alignment and the other breakpoint partner can be inferred from the supplementary alignment. In scenario 2 read pairs, one or both breakpoint partners can be inferred from either the primary alignment or the supplementary alignment. In scenario 3 read pairs, only one, or none of the breakpoint partners can be inferred from the reads.

Identifying breakpoint coordinates

For reads that span a rearrangement breakpoint, it should be possible to infer the breakpoint coordinates at single nucleotide resolution for at least one side of the breakpoint from the BAM alignment information. The way in which we do this depends on whether the 'unmapped' soft clipped sequence is at the beginning or end of the read and whether the relocated region is inverted compared to its original orientation. The interpretation of the location of the soft-clipped sequence becomes a bit more complicated for rearrangements involving inversion, as it is also affected by which of the breakpoint partners is inverted. For example: considering the forward strand of DNA, if the breakpoint partner on the left side of the rearrangement is inverted, the soft-clipped 'unmapped' region will be at the start of the read, whereas if the right-side breakpoint partner is inverted, the soft-clipped 'unmapped' region will be at the end of the read. Which breakpoint partner is inverted also needs to be taken into account when considering the position of any soft-clipping due to Illumina adapter sequences. The way in which breakpoint coordinates are inferred and soft-clipping interpreted, for chromosomal rearrangements involving an inversion, is summarised in Table 4.7.

Table 4.7 Characteristics of reads spanning chromosomal rearrangements involving inversion of one of the relocated segments. The characteristics differ depending on whether the inverted segment is on the 'left' side of the breakpoint fusion or on the right (when considering from the perspective of the forward strand).

	Inverted segment on 'left' of breakpoint fusion	**Inverted segment on 'right' of breakpoint fusion**
Δ read mate coordinates		
Discordant reads	Different chromosomes or ++++	Different chromosomes or ++++
Concordant reads	0	<template length
Breakpoint coordinate		
Forward reads	Start coordinate	Start coordinate + mapped length of read
Reverse reads	Start coordinate	Start coordinate + mapped length of read
Breakpoint soft-clipping		
Forward reads	Beginning of read	End of read
Reverse reads	Beginning of read	End of read
Adapter soft-clipping		
Forward reads	End of read	End of read
Reverse reads	Beginning of read	Beginning of read
SAM flag pairs	(2113, 2225, 2177, 2161) + (97, 145, 161, 181) (2155, 2163, 2179, 2227, 2163) + (83, 99, 147, 163) (113, 177) + (2145, 2209)	(2113, 2225, 2177, 2161) + (97, 145, 161, 81) (2155, 2163, 2179, 2227, 2163) + (83, 99, 147, 163) (65, 129) + (2129, 2193)

4.9.3 Chromosomal rearrangement caller

Armed with an understanding of the characteristic features of paired end reads that span chromosomal rearrangements, a stepwise method was developed that would allow us to call all four possible categories of chromosomal rearrangement. This stepwise process essentially involves five steps as summarised below and involves identifying all possible read pair scenarios in Figures 4.28, 4.29 and 4.30, as well as the identification of both breakpoint partners from all of these reads.

Step 1: Discordant read pair identification

In step 1, the discordant reads are identified, as well as their concordant alternate alignments (if present). This will retrieve all read pairs in Figure 4.28 except h and i.

Step 2: Breakpoint partner identification

In step 2, either one or both breakpoints partners are called, depending on which scenario the read pair belongs to. In Figure 4.28, both breakpoint partners are identified from scenario 1 and scenario 2 read pairs (read pairs a-e) and one breakpoint partner is identified from read pair g. Breakpoint bounds are called from read pairs f.

Step 3: Discordant soft-clip remapping

In step 3, soft-clipped regions are remapped for the discordant reads in which only one breakpoint partner could be called (Figure 4.28, read pair g). The soft-clipped regions of these reads will be <30 bp, or otherwise a concordant supplementary alignment would be present, and so they are at high risk of being mis-mapped if an aligner, e.g. BWA, is used for the remapping. To try to avoid this, information from reads that had already allowed identification of *both* breakpoint partners in step 2 were used. If one of these breakpoint partners had the same coordinates as the discordant soft-clipped read, then the sequence of the other breakpoint partner was compared to the sequence of the soft-clipped region. If there was $> 95\%$ sequence match, then the unknown breakpoint partner in the discordant soft-clipped read could be identified. To try to further avoid mis-mapping, this process was only applied to soft-clipped sequences if they were ≥ 10 bp and breakpoint partners from step 2 had to have been identified from a minimum of 3 sets of paired reads in order to be compared to the sequences of the discordant soft-clipped regions.

Step 4: Concordant read pair identification and concordant soft-clip remapping

In step 4, soft-clipped concordant reads that do not have a discordant alternate alignment, and whose start coordinate matches a breakpoint position already found, are identified (4.28, read pairs h and i). The soft-clipped regions of these concordant reads are then remapped, using the same process used in step 3, except breakpoint partners had to have already been identified from a minimum of 5 sets of paired reads in order to be compared to the sequences of the concordant soft-clipped regions.

Step 5: Chromosomal rearrangement VAF calculation

In step 5, VAFs for both sides of the chromosomal rearrangement are calculated as follows:

$$\text{VAF} = \frac{\text{number of read pairs supporting the chromosomal rearrangement}}{\text{read depth at the breakpoint coordinate for reads with no evidence of chromosomal rearrangement}}$$

VAFs for both sides of the rearrangement are calculated in case one of the breakpoint partners has poor or no panel coverage. The highest of the two VAFs is taken as the final VAF of the chromosomal rearrangement. An output file is produced which details the type of chromosomal rearrangement, the identification of both breakpoint partners, the breakpoint coordinates, the VAF and the number of reads supporting the rearrangement from each of the different read categories in Figure 4.28. Chromosomal rearrangements were ignored if they were called from less than 3 read pairs.

4.9.4 Testing the chromosomal rearrangement caller using simulated data

To test the sensitivity of the chromosomal rearrangement caller, simulated 'samples' were used, each containing a different chromosomal rearrangement from each of the four categories of rearrangement. Samples with a range of different VAFs were created (from 5 to 40%), with each sample containing a different breakpoint position chosen randomly from within the known common breakpoint regions (Appendix C.2.1). The samples were processed using the custom rearrangement caller, with and without steps 3 and 4, to assess how the sensitivity was improved by the addition of the discordant soft-clip remapping and concordant soft-clip remapping steps (steps 3 and 4). If these steps were not included (Figure 4.31a, b), the caller systematically under-estimated the VAF of the rearrangement. Addition of these steps, however, resulted in very good concordance between the known and calculated VAF for each of the samples (Figures 4.31c), although there was some underestimation of the VAF for intrachromosomal rearrangements at VAFs > 10%. The caller was successfully able to detect rearrangements in which only one of the breakpoint partners was covered by our custom panel probes (e.g. because the other breakpoint partner was in a highly repetitive region), although the calculated VAF was slightly underestimated (Figure 4.31, samples highlighted by a '*').

4.9.5 Testing the chromosomal rearrangement caller using patient samples

To further test the sensitivity, as well as specificity, of the chromosomal rearrangement caller, we have obtained 7 DNA samples from patients who each have a known AML-associated chromosomal rearrangement (samples kindly provided by Dr Peter Valk, Erasmus University Medical Centre). These 7 samples include all of the AML-translocations covered by our panel, except for t(16;16) and t(15;17). We plan to process these samples in the same way as the UKCTOCS samples and then determine if our chromosomal rearrangement caller can detect the known rearrangements. This work is still ongoing. Pending validation of our chromosomal rearrangement caller on samples with known rearrangements, we also used Manta[195] to call rearrangements in the UKCTOCS samples.

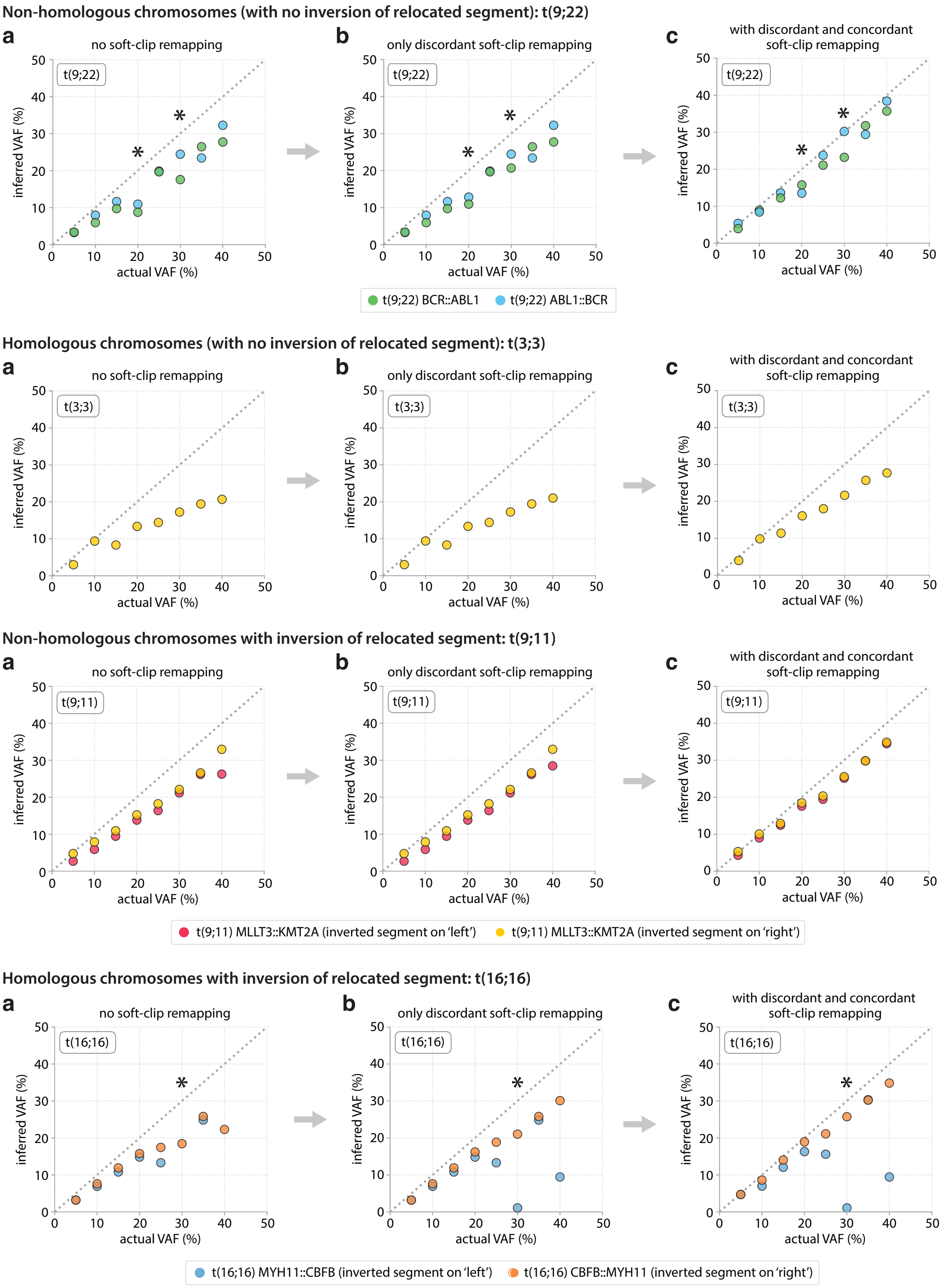


Figure 4.31 Calling chromosomal rearrangements using simulated samples in each of the four rearrangement categories, with sequential improvement in VAF concordance with the addition of discordant soft-clip remapping and concordant soft-clip remapping. **a.** Concordance between actual and inferred VAFs when the caller did not perform any remapping of soft-clipped regions. **b.** Concordance between actual and inferred VAFs when the caller remapped soft-clipped regions of discordant reads, but not concordant reads. **c.** Concordance between actual and inferred VAFs when the caller remapped soft-clipped regions of both discordant and concordant reads. Samples in which one of the breakpoint partners was not covered by the custom panel are highlighted with an '*'.

4.10 Discussion

Wanting to trace the clonal evolution to AML through time, using longitudinal pre-AML blood samples, meant we needed to be able to detect a comprehensive array of genomic changes and we needed to be able to detect them when they were in just a small number of cells. In this chapter we described the development of an integrated approach, using duplex sequencing and *in silico* noise correction methods to allow reliable detection of gene mutations, chromosomal rearrangements and mCAs when they are present at very low cell fraction. Optimisation of this method is still ongoing and there are several aspects that can be further improved.

Whilst the *in silico* noise correction method allows us to identify real variants from amongst errors, the lowest VAF at which a variant could be detected is, for some positions, determined by the position's overall error-rate. The use of duplex sequencing helps to reduce this error-rate, but whilst the theoretical error rate of duplex sequencing is often quoted as $< 10^{-9}$ error per bp (which would theoretically mean we could detect VAFs nearly as low as this, if we started with sufficient DNA molecules), we and others have found that the error rate is actually significantly higher[14,141]. Our overall error-rate across our panel was 2.6×10^{-5}, although some positions had error rates as high as $\sim 10^{-2}$. The source of these errors is thought to be library preparation artefacts e.g. due to sonication, end-repair and mapping errors[14,141,188] which can all slip through the duplex error correction. A recently published whole-genome duplex error-corrected method, called NanoSeq by Abascal et al, provided a detailed analysis of these errors and described ways in which they could be reduced[141]. This enabled them to reduce their duplex error rate from $\sim 2 \times 10^{-7}$ to $< 5 \times 10^{-9}$. Three of the key adaptations they incorporated were i) avoiding end-repair, ii) blocking nick extension and iii) reducing mapping errors.

'End-repair' is a necessary step during library preparation after the DNA has been fragmented (using either sonication or fragmentation enzymes) and can result in single-stranded DNA damage being converted into double stranded errors. Abascal et al[141] managed to avoid this by using specific fragmentation enzymes which create blunt ends, thus avoiding the need for any end-repair. A disadvantage of this approach is the incomplete genome coverage of the endonucleases and so an alternative approach they suggested was to use sonication combined with digestion of the overhanging ends using Mung Bean exonuclease. This approach is good for applications requiring whole-genome coverage, but comes at the cost of poor library yields: 10-50× lower than using fragmentation enzymes[141]. Because we have limited amounts of DNA available for our UKCTOCS samples, we need to maximise our yield as much as possible, which is why we use enzymatic fragmentation rather than sonication. In our library preparation we currently use proprietary fragmentation enzymes provided by Twist Biosciences. Changing these fragmentation enzymes for a selection of blunt-end-creating endonucleases (each covering a range of genomic regions), could be an option to attempt to reduce errors introduced during end-repair.

‘Nick extension’ involves A-tailing of DNA fragments prior to adapter ligation and involves a DNA polymerase and deoxyadenosine triphosphates (dATPs). Abascal et al[141] observed increased levels of C>A, G>A and T>A base changes at restriction enzyme sites and reasoned that when the double-stranded DNA was nicked by the restriction enzymes (as an intermediate step in the double-strand cleave), 3’-to-5’-exonuclease or pyrophosphylation of the dNTP 3’ of the nick occurs and a dATP is incorporated during the A-tailing step, resulting in a G>A, C>A or T>A error. To resolve this issue they replaced dATP with a mixture of dATP and dideoxynucleotides triphosphates (ddNTPs), ddCTP, ddGTP, ddTTP, during the A-tailing step. In contrast to deoxynucleotides triphosphates (dNTPs), dideoxynucleotides triphosphates (ddNTPs) lack 3’-OH groups and so are are unable to form phosphodiester bonds with the next nucleotide. This means that when the DNA polymerase attempts to extend at the internal nick sites, the incorporation of a ddNTP results in a strand that is unamplifiable. Further work is needed to determine if this ‘nick extension’ is introducing a significant amount of errors in to our workflow, because although our G>A error rates are high, our C>A and T>A error rates are comparatively low (Figure 4.9 and 4.19). Nonetheless, incorporating ddNTPs into our A-tailing step is an adaptation we could consider.

Abascal et al[141] observed that unambiguous mapping could result in miscalculation of error rates, particularly in highly repetitive regions. To avoid this, they discarded reads in which the minimum difference between the primary alignment score and the secondary alignment score was <50 and discarded variant calls within 8 bp of the ends of reads, where unreliable mapping was more likely[141]. They also excluded read pairs that were ‘improperly paired’. In our workflow, we already excluded variants whose mean position in the read was <8 bp and only included read pairs that were flagged as being ‘mapped in a proper pair’ by BWA. With our targeted exonic SNV/ indel panel, mismapping due to highly repetitive regions is unlikely to be as much of a problem as in whole genome duplex sequencing, but, nonetheless, incorporating this step may help to reduce our error rates, as it may explain why we observed some positions with error rates as high as 10^{-1}. Further work is required to further explore the cause of errors at these positions.

There are also ways in which our *in silico* noise correction method can be improved. Our approach is similar to the method used in Shearwater[200], a variant caller which estimates position-specific error profiles using a beta-binomial distribution, and uses prior information about mutational hotspots from the COSMIC database. Shearwater estimates the beta-binomial dispersion parameter using the method-of-moment estimator whereas we chose to use the method-of-moment estimate to initialise a maximum likelihood approach for estimating the dispersion, as we found this yielded more accurate estimates, particularly for positions with high dispersion (Appendix C.1.1). Our approach uses an iterative approach at COSMIC[166] sites, whereas Shearwater takes into account the prior probability that a variant exists from the distribution of observed somatic mutations at the site in COSMIC[166]. Even with the iterative approach at COSMIC sites, our method risks missing real variants if there are two or more samples at the position with similarly high VAFs (Figure 4.21), due to the fitting of a falsely over-dispersed beta-binomial distribution. It is possible that Shearwater may also miss these variants, but incorporating aspects of the Shearwater approach for COSMIC sites is something that we may adopt. At COSMIC sites it seems sensible to adjust the *p*-value threshold according to how frequently the mutation is observed in COSMIC. It would also be prudent to set a prior on the fitted dispersion parameter, to try to avoid real variants being interpreted as errors.

We have been fairly conservative with our variant calling approach thus far, choosing to minimise our false discovery rate as much as possible (to <5%). However this likely comes at the cost of increased false negatives rates and so it is possible that we have inadvertently filtered out some real variants. Incorporating the above-mentioned error-reducing adaptations to our library preparation and further optimising our *in silico* noise correction method should allow us to decrease our position-specific error rates even further and allow us to identify any additional real variants that we might have otherwise missed.

Overall, our integrated approach combines the genomic breadth required to detect an array of key pre-AML associated genomic changes, with the depth and accuracy of error-correcting techniques to detect the mutations when they are present in only a small number of cells. Considering the breadth and depth that the approach achieves, its current cost of $\sim$£225 per sample (including sequencing) is not cost-prohibitive. Our approach enables us to comprehensively trace the clonal evolution of mutations over time in the years preceding AML diagnosis, which should allow us to disentangle the whole evolutionary process, from acquisition of the first driver mutation through to AML diagnosis.

5

Tracing the evolution to AML using longitudinal pre-diagnosis blood samples

5.1 Introduction

For the majority of individuals with clonal haematopoiesis, the presence of pre-leukaemic mutations will be of little or no consequence. However, if further mutations are acquired and clonal expansion occurs, then clonal haematopoiesis can progress to acute myeloid leukaemia (AML)[17,18,24]. For individuals with a somatic mutation ≥2% VAF, progression to AML (or other haematological malignancy) occurs, on average, at a rate of ∼ 0.5-1% per year[29]. AML is an aggressive blood cancer, which despite recent advances in treatments, still has an overall 5 year survival of <30%[201]. If we can understand how and why clonal haematopoiesis progresses in some individuals and not others then we may be able to intervene early and stop AML in its tracks before it fully develops.

The genomic landscape at the time of AML diagnosis has been well-characterised, with an average of 5 ± 3 (SD) mutations in recurrently mutated genes being found at the time of diagnosis[20,21]. Based on known pathways and functional analyses, genes recurrently mutated in AML are typically grouped into functional-biologic categories with shared co-occurrences and exclusive dissociations: NPM1, DNA methylation (e.g. DNMT3A, TET2), chromatin modifiers (e.g. ASXL1, EZH2), myeloid transcription factors (e.g. RUNX1, CEBPA), tumour-suppressors (e.g. TP53, WT1), spliceosome-complex (e.g. SRSF2, SF3B1), cohesin-complex (e.g. STAG2, RAD21) and activated cell signalling-pathway (e.g. FLT3, KRAS) gene mutations[21]. The genomic landscape of clonal haematopoiesis is also well-described[3,4,7,9,75,202], including an understanding of the fitness effects and mutations rates of recurrently mutated genes and mCAs[93] (Chapter 2 and Chapter 3). How the mutational landscape differs in individuals who progress to AML was explored in recent studies, using blood samples collected from up to 3 timepoints pre-diagnosis[37–39]. These studies highlighted key features associated with increased AML risk, which were detectable in the blood ∼10 years prior to AML diagnosis. Pre-AML cases showed enrichment for mutations in IDH1, IDH2, TP53, DNMT3A, TET2 and spliceosome genes. VAFs tended to be higher in pre-AML cases, with ∼40% having a mutation at >10% VAF, compared to only ∼4% of controls[37]. Clonal complexity was also greater in pre-AML cases, with ≥1 mutations detectable in nearly 50% of cases compared to only ∼5% of controls, the majority of whom had 0 mutations ($\gtrsim$1% VAF)[38].

These studies captured 'snapshots' of the pre-leukaemic evolutionary process, but what we don't have is a complete view of the whole evolutionary process from acquisition of the first driver mutation through to AML diagnosis. Several key questions remain unanswered (Figure 5.1): At what age does the initiating driver mutation occur? How long does it take to acquire the 2nd mutation? Are the fitness effects conferred by specific mutations predictable from person-to-person and how do fitness effects change with additive mutations? What is the pattern of clonal evolution through to AML? Knowing the answers to these questions is important if we want to develop risk stratification methods for individuals with clonal haematopoiesis and identify individuals that may benefit from early intervention studies.

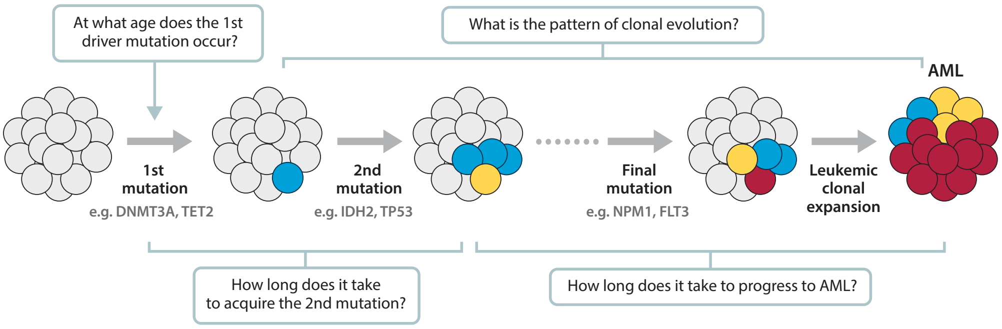

At what age does the 1st driver mutation occur?
What is the pattern of clonal evolution?
AML
1st mutation
e.g. DNMT3A, TET2
2nd mutation
e.g. IDH2, TP53
Final mutation
e.g. NPM1, FLT3
Leukemic clonal expansion
How long does it take to acquire the 2nd mutation?
How long does it take to progress to AML?

Figure 5.1 Pre-leukaemic evolution: Unanswered questions.

To try to answer these questions we identified 50 women from the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) who were cancer-free at enrolment but subsequently developed AML during the >12 years follow-up, at an average age of 71. These women had provided yearly blood samples during their involvement in the trial, some at up to 11 timepoints (average 5 timepoints). We obtained all of their pre-AML yearly blood samples as well as blood samples from 50 age and timepoint matched controls (see Chapter 4 Section 4.1.1).

In this chapter we describe our preliminary findings from the analysis of this unique set of longitudinal blood samples. We find that there are four main evolutionary patterns in the years preceding AML diagnosis: linear evolution, evolution with clonal interference, static evolution and late evolution. We calculate the age of acquisition of the first and second mutations and whilst the initial driver mutation is often acquired early in life, there are some very fit 'uber drivers' which appear to occur as the initial event just $\sim$ 4 years pre-diagnosis. The 20 variants, which we identified as 'highly fit' in clonal haematopoiesis[93] (Chapter 2 Section 2.5.1) are significantly enriched in pre-AML cases compared to controls and we determine how the fitness effects change with the acquisition of subsequent mutations. These findings reveal key insights into the evolutionary dynamics of clones in the years preceding AML development.

5.2 Reconstruction of clonal evolutionary histories

The longitudinal blood samples were processed and sequenced using our custom comprehensive sequencing panel, using duplex error-corrected sequencing and *in silico* noise correction methods, as described in Chapter 4. Having samples from multiple timepoints allowed us to visualise clear trajectories of the clonal evolution in the years pre-AML diagnosis, as well as in the controls (Figure 5.2 and Appendix D.1). Clonal composition could be inferred in many individuals by using VAF information and growth trajectory dynamics. For example, if the combined VAF of 2 mutations was >50% in a later timepoint, then they could be inferred to have arisen in the same clone. Similarly, if two mutations followed the same growth trajectory, this also increased the likelihood of them being present in the same clone.

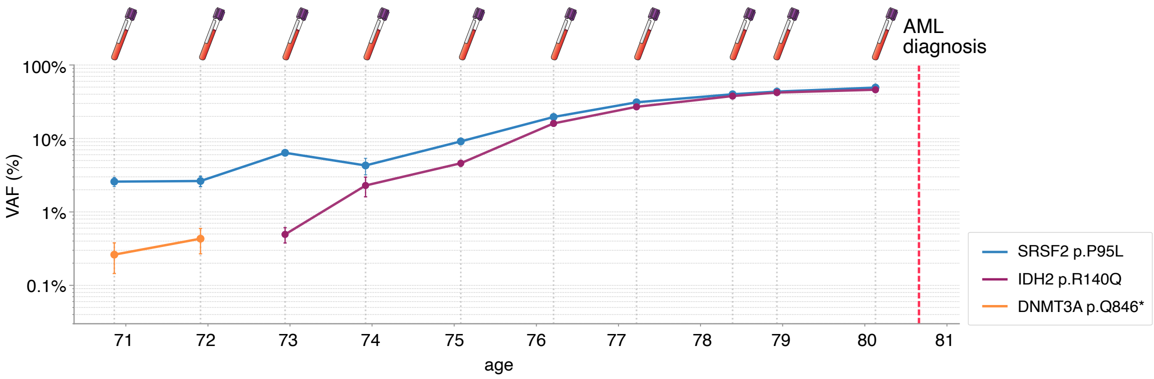


Figure 5.2 Pre-AML trajectory of clonal evolution. In this example, an SRSF2 P95L and IDH2 R140Q mutation are both detectable at nearly 50% VAF ∼ 6 months before AML diagnosis. Their summed VAFs are >50% and so we can infer that they must be present in the same clone. Tracing their trajectories back in time we can see the order in which they co-occurred and can estimate their age at acquisition (see Section 5.3).

5.2.1 Clonal complexity and clonal relationships of driver mutations

Consistent with previous studies, the number of mutations detected, at any VAF, was significantly greater in pre-AML cases than in controls, as far back as 8 years pre-diagnosis (Figure 5.3a) with an average of 2.6 mutations in pre-AML cases and 1.3 mutations in controls ($p = 0.011$, unpaired T-test). This was particularly prominent when considering mutations at ≥10% VAF, which was significantly greater in pre-AML cases even >10 years pre-diagnosis, with an average of 0.8 mutations >10% VAF in pre-AML cases and 0.2 in controls ($p = 0.027$, unpaired T-test) (Figure 5.3b). The closer the sample was taken to diagnosis, the more significant the difference between pre-AML cases and controls, with 1.7 mutations >10% VAF in pre-AML cases 0-2 years pre-diagnosis, compared to 0.15 in controls ($p = 1.3 \times 10^{-5}$) (Figure 5.3b).

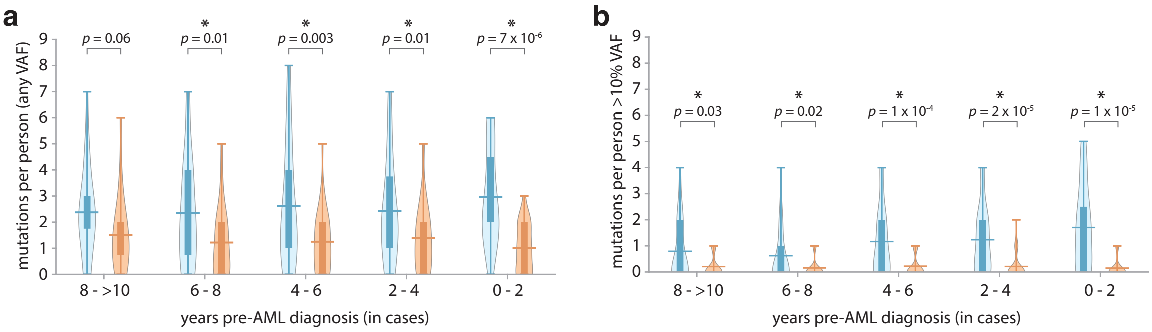


Figure 5.3 Total number of mutations in pre-AML cases and controls. Pre-AML cases are represented in blue and controls are represented in orange. The wide horizonal line in the violin plots represents the mean of the distribution and the thick bar represents the interquartile range. The width of the violin plot represents the proportion of data located there. Statistical significance between pre-AML cases and controls at each timepoint was tested using a two-sided unpaired t-test (p-values shown). **a.** Total number of mutations at any VAF. **b.** Total number of mutations at >10% VAF.

Multiple mutations within the same gene were detected in a number of samples and having the ability to reconstruct clonal composition, from VAF measurements and clonal trajectory dynamics, meant we could determine the clonal relationships for many of these mutations. Multiple TET2 mutations were particularly common in pre-AML cases, being observed in 20% of pre-AML cases, compared to only 4% of controls (Figure 5.4). These appeared to be in separate clones in the controls, but but appeared to co-occur in the same clone in all the pre-AML cases (as either 2 somatic, 1 somatic/ 1 germline or 2 germline), consistent with the well-described presence of biallelic TET2 mutations in myeloid neoplasms[203] (Appendix Figure D.1). One pre-AML case had 3 WT1 mutations and at least 2 of these were present within the same clone, consistent with previous work describing biallelic involvement of WT1 in AML[204].

The presence of two or more DNMT3A mutations was recently reported to be associated with a significant increased odds ratio of AML (odds ratio 12.6)[76] and whilst we found that multiple DNMT3A mutations were common in our cohort, they appeared to be less common in pre-AML cases than in controls (8% of pre-AML cases and 16% of controls) (Figure 5.4). These multiple DNMT3A mutations appeared to be present in the same clone in ∼1/2 of the pre-AML cases and ∼ 1/3 of controls (Appendix Figure D.1). Whilst biallelic DNMT3A mutations are found in >60% of patients with T cell acute lymphoblastic leukaemia (T-ALL), biallelic DNMT3A mutations in myeloid neoplasms are much less common, which is thought to be due to the requirement for DNMT3A activity for myeloid lineage choice[205]. The lack of enrichment for biallelic DNMT3A mutations in our pre-AML cases is consistent with this.

Previous bulk sequencing and single-cell studies have reported a tendency to mutual exclusivity for DNMT3A, TP53, TET2 and ASXL1, which is presumed to be due to functional redundancy[21,38,206]. Whilst our numbers are too small to draw firm conclusions for most of these relationships, we did not observe clear evidence for this for DNMT3A, which was detected in the same clone as TET2, ASXL1 and TP53 in at least 3 separate pre-AML cases (Figure, 5.4, Appendix Figure D.1).

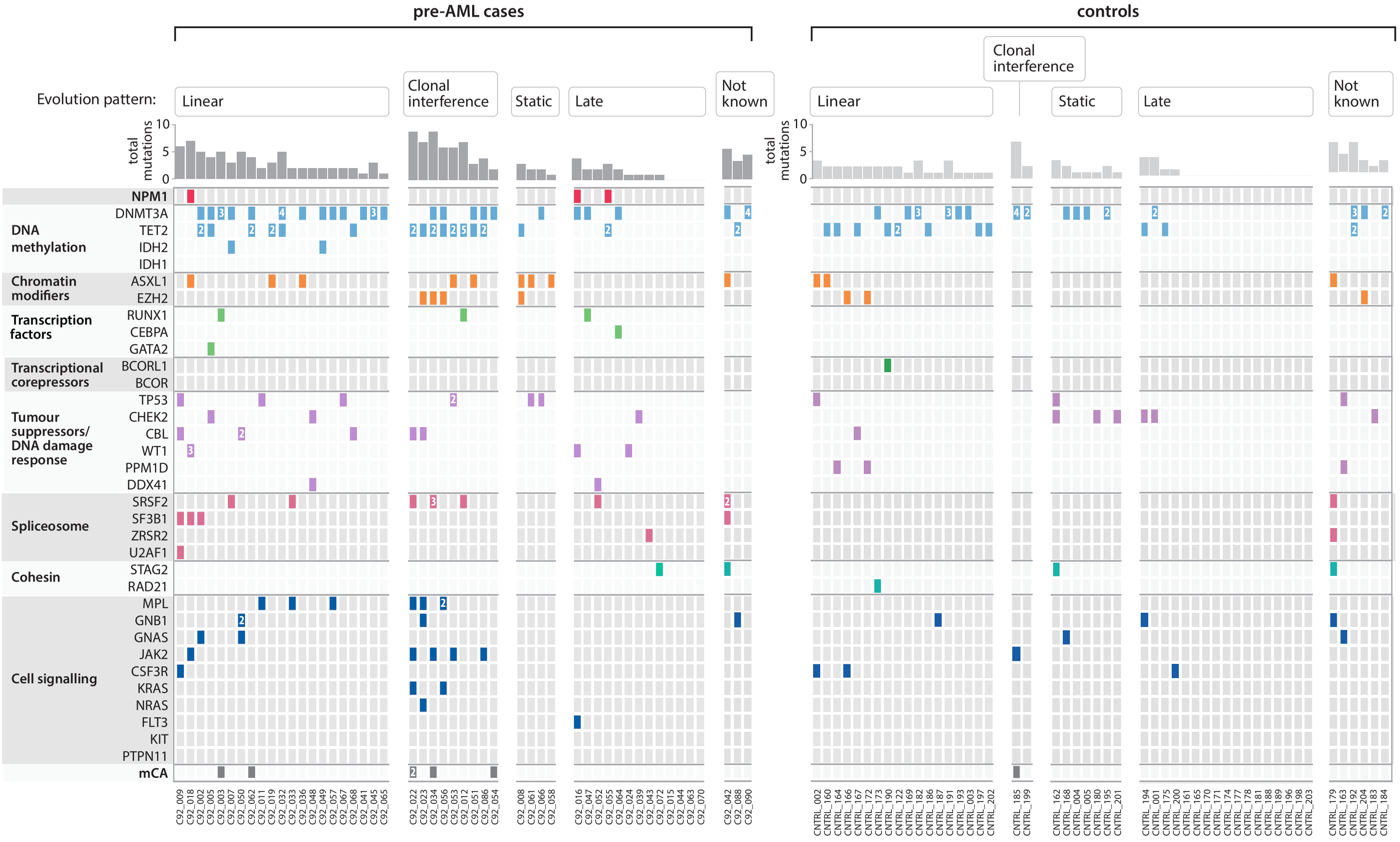


Figure 5.4 Mutations detected in pre-AML cases and controls. Pre-AML and controls are subcategorised by their observed pattern of evolution (as discussed in section 5.2.2). If more than one mutation was detected in a particular gene, this is indicated by the number inside the grid box.

5.2.2 Evolutionary patterns of AML development

By reconstructing the clonal evolution trajectories for all 50 pre-AML and control samples, it was visually evident that there were four main evolution patterns: i) linear evolution, ii) evolution with clonal interference, iii) static evolution and iv) late evolution.

Linear evolution

The linear evolution pattern was most common and was observed in 21 out of 50 pre-AML cases and 18 out of 50 controls. In this evolution scenario, successive mutations were acquired in a step-wise manner in a single dominant clone, with clear evidence of unopposed exponential growth across all sample timepoints.

Some samples that showed linear evolution can be see in Figure 5.5. For example, in Figure 5.5a, a quadruple-mutant clone containing an SF3B1, DNMT3A and 2 different TET2 mutations was already detectable 7 years before AML diagnosis. Mutation acquisition order can be inferred from the VAFs and reveals that the DNMT3A mutation occurred first, followed by SF3B1 then TET2, followed by another TET2 mutation. This quadruple-mutant clone grew unopposed across all the timepoints, reaching nearly 40% VAF $\sim$ 18 months before diagnosis (Figure 5.5a). In Figure 5.5b, a double-mutant clone containing a JAK2 and SF3B1 mutation was detectable 8 years before AML diagnosis. This clone can be seen growing exponentially, until it acquired an NPM1 frameshift mutation 2 years pre-diagnosis followed by 2 WT1 mutations $\sim$ 1 year pre-diagnosis, which resulted in a dramatic increase in growth rate with considerable clonal expansion.

Evolution with clonal interference

The clonal interference pattern was observed in 9 out of 50 pre-AML cases, but only 2 out of 50 controls. In this evolution scenario there are multiple clones present, with clear evidence of clonal competition between them.

Some samples that showed evolution with clonal interference are shown in Figure 5.6. For example, in Figure 5.6a, an SRSF2 P95H mutation has already clonally expanded to 45% VAF (90% cell fraction) 8 years pre-AML diagnosis. This clone then acquires a JAK2 V617F mutation, which confers a significant growth advantage to the SRSF2/JAK2 clone, but this starts to be outcompeted by the acquisition of a 19p CN-LOH event in a different SRSF2 subclone. The SRSF2/JAK2 clone acquires an additional 9p CN-LOH event, however, which shows signs of starting to become the dominant clone again $\sim$ 2 years before AML diagnosis. In Figure 5.6b multiple subclones, which have arisen on the background of an SRSF2 mutant clone, are detectable 7-8 years pre-AML diagnosis, with the TET2 Q917X subclone (light green) having a significant fitness advantage over the others. The TET2 N1266S subclone (dark green) then gains a competitive advantage, however, and outcompetes the

TET2 Q917X clone (light green), becoming the dominant clone in the year prior to AML diagnosis. No new mutations were detected to explain the TET2 N1266S subclone's competitive advantage.

In two of the samples that showed a clonal interference evolution pattern (Figure 5.6b and c), the clonal competition appears to occur within a constrained fraction of the total number of cells. In Figure 5.6b, the TET2 Q917X clone has seemingly reached its maximum size at age $\sim$ 66, seemingly resulting in a decrease in the size of the other SRSF2 subclones. Yet the 'wild-type' haematopoietic stem cells still make up 34% of the cells at this point and the SRSF2 subclones would be expected to outcompete these before they were reduced in size by the TET2 Q917X subclone. Similarly, in Figure 5.6c, as the MPL clone starts to increase in size, the KRAS clone decreases in size, but this clonal competition is constrained within $\sim$40% of the cells across all the timepoints.

Static evolution

The static evolution pattern was observed in 4 out of 50 pre-AML cases and 7 out of 50 controls. In this evolution scenario, the clones have clearly undergone clonal expansion, in order to reach a level where they were detectable, but showed no growth across a number of sequential timepoints.

Some samples that showed static evolution can be seen in Figure 5.7. For example, in Figure 5.7a, a TP53 mutant clone, which has clonally expanded to $\sim$ 9.5% VAF by the age of 64, shows no growth at all across 3 timepoints spanning a 1 year period. Similarly, an ASXL1 mutant clone in the same individual, which had clonally expanded to $\sim$1% VAF by the age of 64, also shows no growth over the same period. In Figure 5.7d, a TET2 mutant clone can be seen growing exponentially during the first $\sim$2 years of samples (4-5 timepoints), from a VAF of 8% to 27%, but the clone then seems to stop growing and remains the same size for the next $\sim$ 2.5 years (5 timepoints), $\sim$ 5 years before AML diagnosis. No obvious clonal competitor was detected, from within our targeted panel, in any of the samples that showed a static evolution pattern.

Late evolution

The late evolution pattern was observed in 13 out of 50 pre-AML cases. In this evolution scenario there were either no mutations detectable, or mutations were only detected in the 1-2 sample(s) closest to the time of AML diagnosis.

Some samples that showed late evolution can be seen in Figure 5.8. For example, in Figure 5.8a, no mutations are detected until $\sim$ 18 months before diagnosis, when an NPM1 frameshift mutation becomes detectable, followed very quickly by a WT1 mutation with significant clonal expansion to $\sim$8% VAF over the course of 1 year. An additional FLT3-ITD mutation then becomes detectable $\sim$ 3 months before diagnosis. In Figure 5.8c, no mutations were detected until $\sim$ 3 months before diagnosis, when a WT1 frameshift mutation becomes detectable at 13% VAF, having been undetectable just 1 year previously.

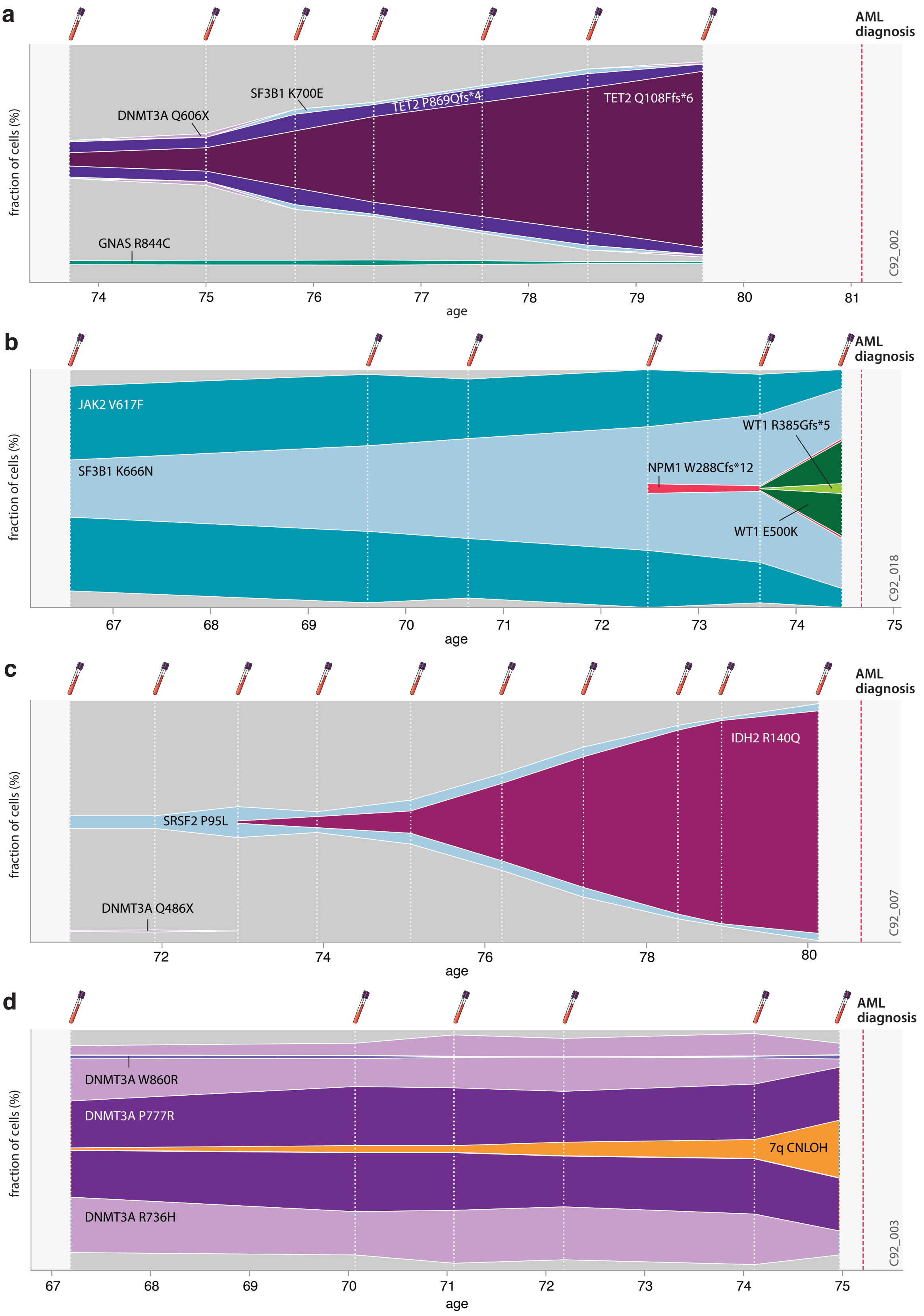


Figure 5.5 Examples of samples showing a linear evolution pattern. Vertical dashed lines indicate the timings of the blood samples. Clonal relationships were determined from the clonal trajectories, which can be found in Appendix D.1.

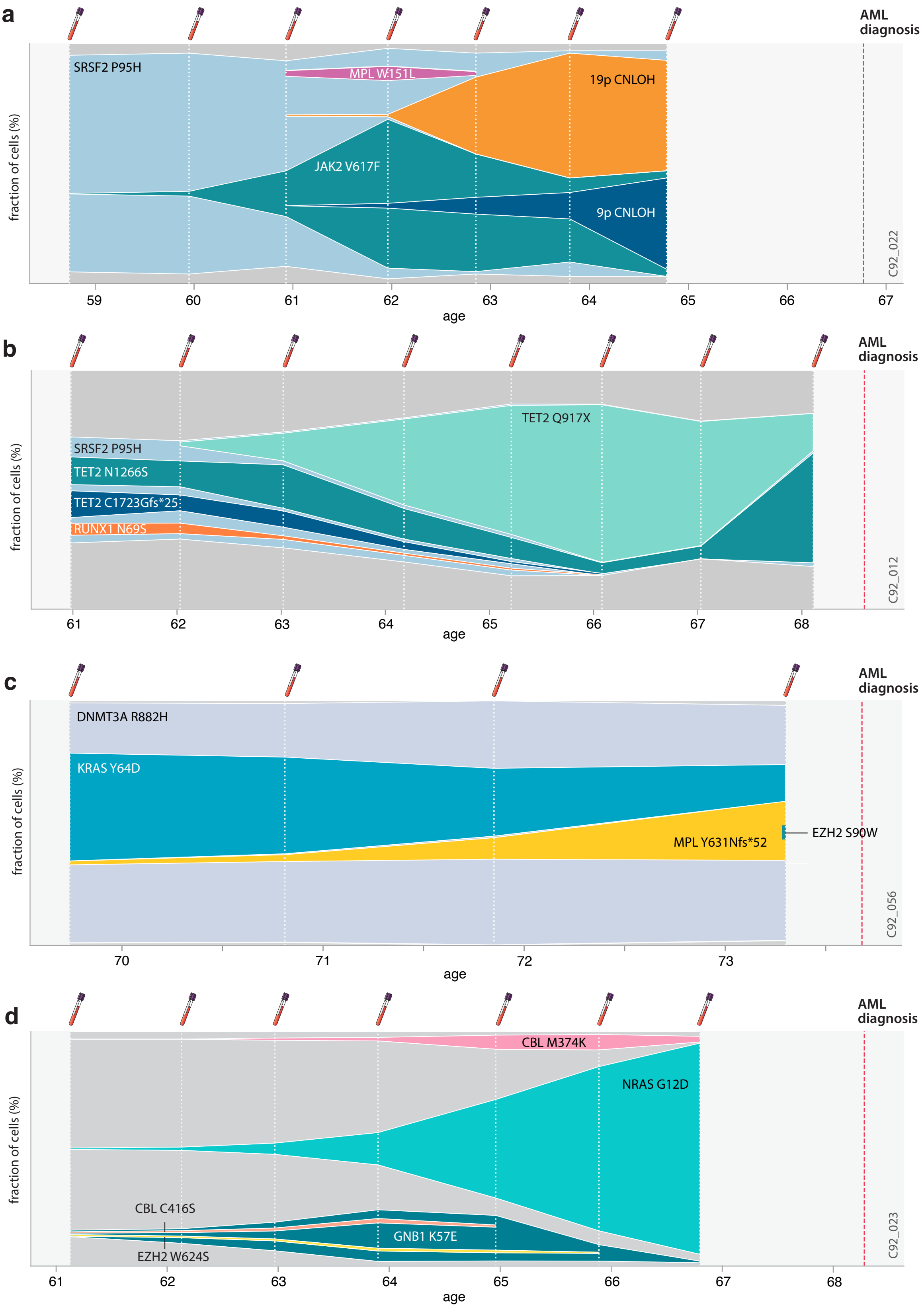


Figure 5.6 Examples of samples showing a clonal intereference evolution pattern. Vertical dashed lines indicate the timings of the blood samples. Clonal relationships were determined from the clonal trajectories, which can be found in Appendix D.1.

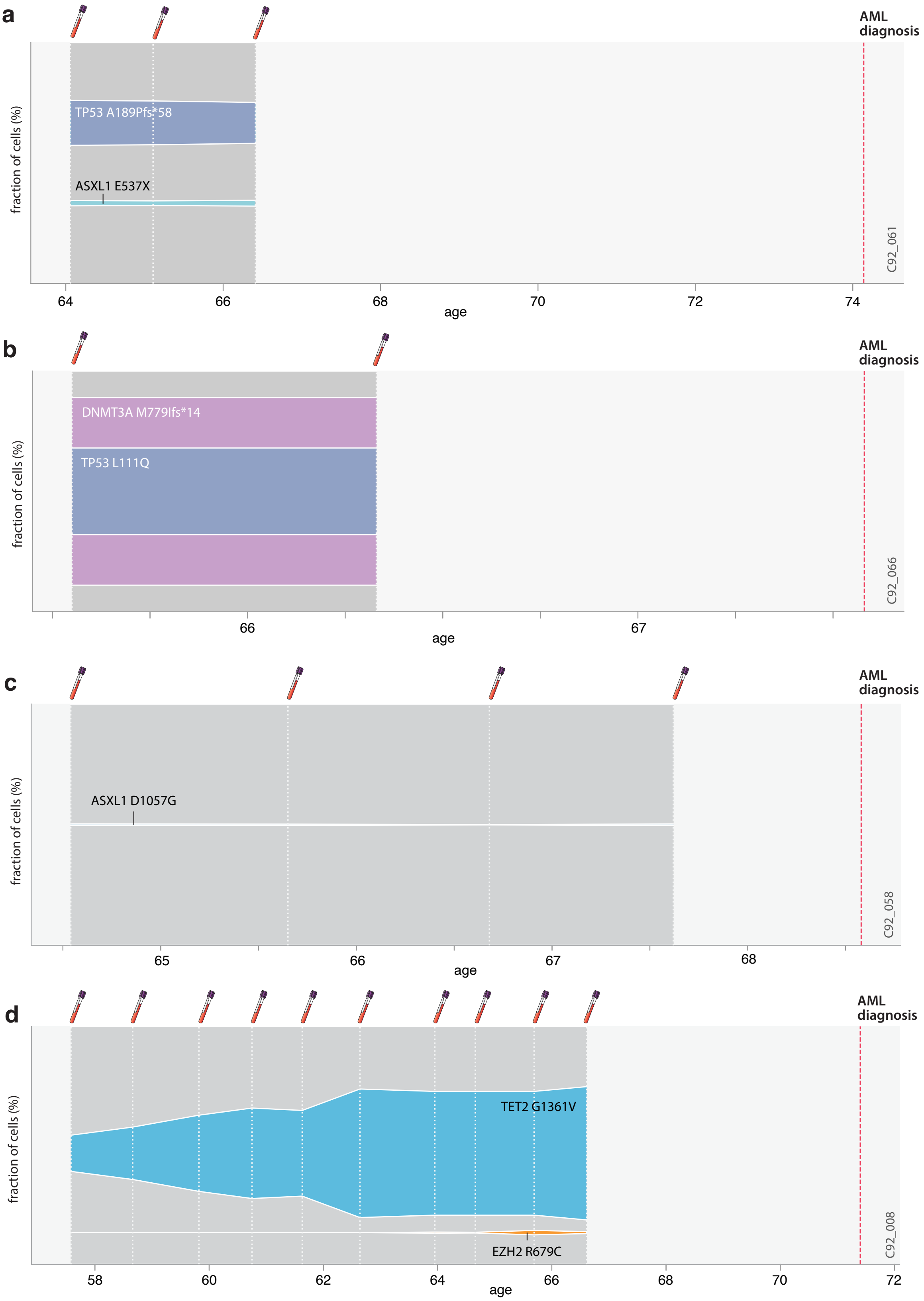


Figure 5.7 Examples of samples showing a static evolution pattern. Vertical dashed lines indicate the timings of the blood samples. Clonal relationships were determined from the clonal trajectories, which can be found in Appendix D.1.

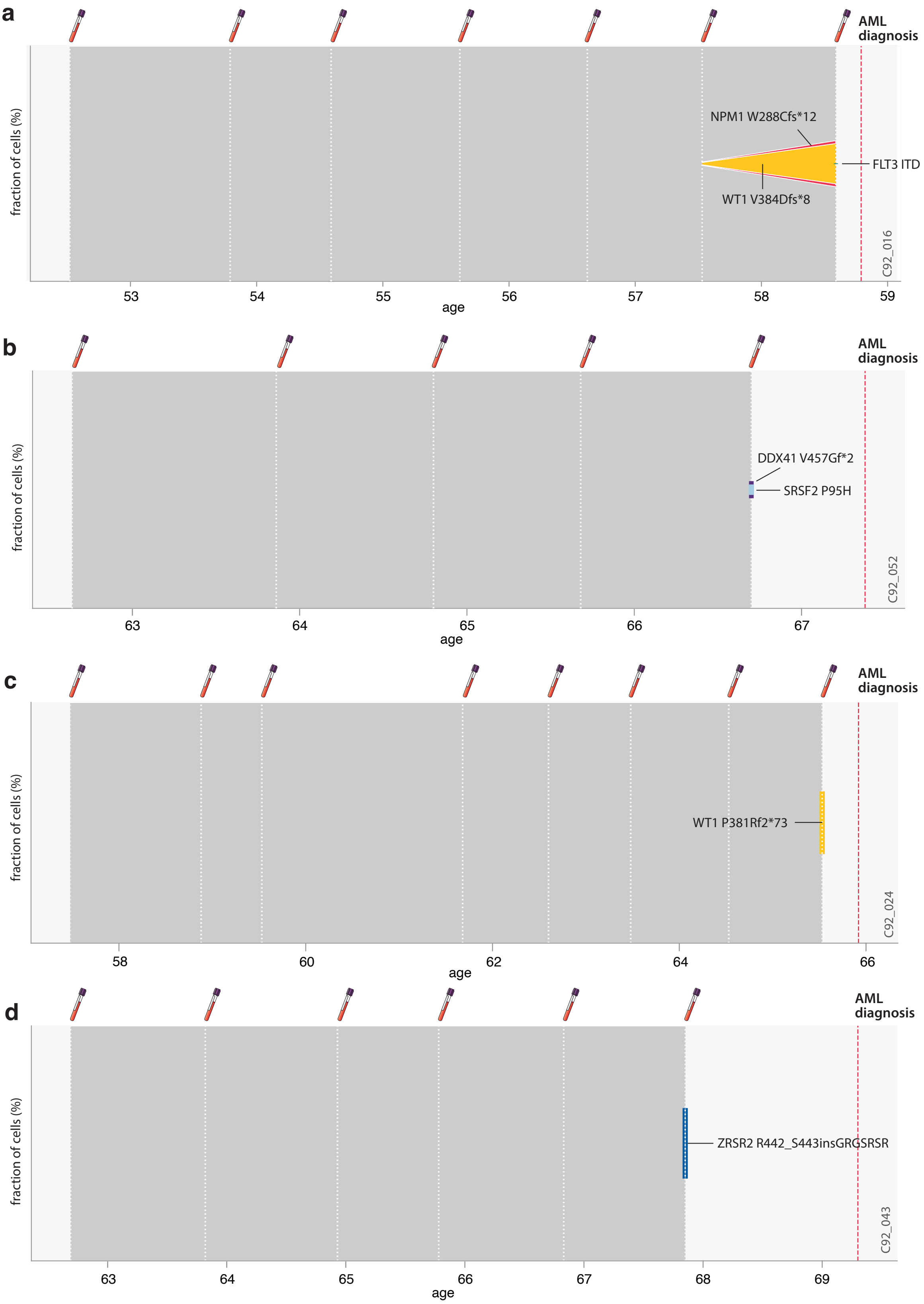


Figure 5.8 Examples of samples showing a late evolution pattern. Vertical dashed lines indicate the timings of the blood samples. Clonal relationships were determined from the clonal trajectories, which can be found in Appendix D.1.

We wondered whether different clonal evolution patterns were characteristic of specific gene mutations. The sample numbers are too small to draw firm conclusions, but we did not see any static evolution occurring in individuals with mutations in spliceosome or cell signalling genes (Figure 5.4).

5.3 Age at acquisition and fitness effects of initiating driver mutations

By reconstructing the clonal composition of the samples that showed linear and late evolution, it is possible to estimate the timings of the key steps in the evolution to AML for these two evolutionary scenarios. Fitness effects for the initial driver mutations can also be estimated, as well as the effect on fitness of acquisition of additional mutations. Estimating these parameters is more complex for the samples with a clonal interference pattern of evolution, due to the variation in growth rates caused by clonal competition. It is also difficult to estimate these parameters for the samples that showed a static pattern of evolution, due to uncertainties regarding the growth dynamics that preceded the halt in growth, or at what age it occurred.

5.3.1 Linear evolution pattern

For samples that show a linear evolution pattern, the age at acquisition of the first and second driver mutations can be determined by effectively extrapolating the exponential trajectories of the first two mutations back in time. At age T, the number of cells containing the first mutation alone (n_1), acquired at age t_1, is expected to be:

$$n_1 = \frac{e^{s_1(T-t_1)} - 1}{s_1} \tag{5.1}$$

where s_1 is the fitness effect of the clone containing just the first mutation

and the number of cells containing the second mutation (n_2), acquired at age t_2, is:

$$n_2 = \frac{e^{s_2(T-t_2)} - 1}{s_2} \tag{5.2}$$

where s_2 is the fitness effect of the clone containing both the first and the second mutation

Before the second mutation is acquired (i.e. $T < t_2$), the total number of cells containing the first mutation is simply n_1. However, after the second mutation is acquired (i.e. $T \geq t_2$) the number of cells containing the first mutation also includes the cells that acquired the second mutation. Therefore when $T \geq t_2$ the number of cells containing the first mutation is $n_1 + n_2$.

The VAF for the cells containing the first mutation (VAF$_1$) can therefore be calculated as:

$$\text{VAF}_1 \text{ (when } T < t_2) = \frac{n_1}{2(N+n_1)} \qquad \text{VAF}_1 \text{ (when } T \geq t_2) = \frac{n_1+n_2}{2(N+n_1+n_2)} \tag{5.3}$$

where N is the total number of wild-type haematopoietic stem cells and the VAF for the cells containing the second mutation (VAF_2) can be calculated as:

$$\text{VAF}_2 \text{ (when } T \geq t_2) = \frac{n_2}{2(N + n_1 + n_2)} \tag{5.4}$$

A maximum likelihood approach was used, minimising the L2 norm between the observed and expected VAFs, to extrapolate the trajectories. These inferred trajectories fit the observed VAFs very well, consistent with exponential growth, and enabled us to determine the age at acquisition of the first and second mutations (t_1 and t_2) and the fitness effect of the single-mutant clone (s_1) and double mutant clone (s_2) (e.g. Figure 5.9). These estimates naturally depend on the number used for N (total number of wild-type haematopoietic stem cells), a number which is not definitively known. However, work by us[93] and others[92,124] has estimated it to be $\sim$ 20,000 to 200,000 and so we used an estimate of $N = 100,000$ in our maximum likelihood estimations. Increasing the value used for N would result in greater estimates for the fitness effects (s_1 and s_2) and/ or younger estimates for the acquisition ages (t_1 and t_2).

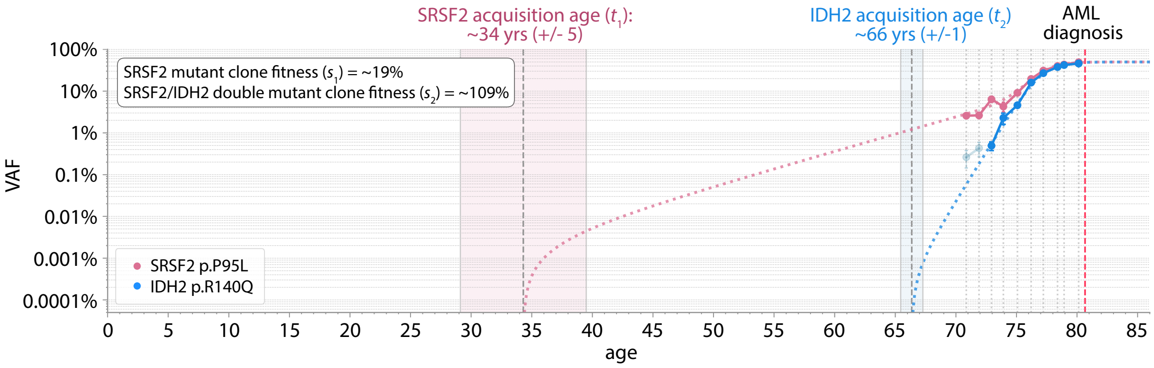


Figure 5.9 Estimation of acquisition age and fitness effects for mutations showing a linear evolution pattern. Solid colour datapoints indicate the measured VAF from the blood samples at each timepoint (each timepoint indicated by a vertical grey dashed line). The dashed coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches, which allow estimation of the acquisition age for the first (SRSF2) and second (IDH2) mutation, as well as the fitness effect (s) of the single-mutant (SRSF2) and double-mutant (SRSF2/IDH2) clone. The error measurement shown for the acquisition ages is $\pm 1/s$. Any other mutations present in the sample are shown as faded out datapoints.

Linear evolution: Age at acquisition of driver mutations

By applying this method to all the samples that showed a linear evolution pattern (Appendix D.2), we find that the average age at acquisition of the first driver mutation is $\sim$ 22 years old (range 0 - 55 years) in the pre-AML cases and $\sim$ 19 years old (range 0 - 62 years) in the controls (not statistically significantly different, $p = 0.30$, two-sample Kolmogorov-Smirnov test) (Figure 5.10, left plot).

In the pre-AML cases, it takes on average $\sim$ 21 years (range 0 - 42 years) from the first mutation to acquire the second mutation (Figure 5.10, middle plot) and then the time between acquiring the second mutation and diagnosis of AML is $\sim$ 30 years (range 12 - 60 years) (Figure 5.10, right plot).

Although our sample numbers are too small to determine clear patterns regarding typical acquisition ages for classes of gene mutations, the age at acquisition of DNA methylation (e.g. DNMT3A, TET2) and tumour suppressor (e.g. TP53, CHEK2) gene mutations, as the first driver mutation, were spread across the range of acquisition ages in both pre-AML cases and controls (Figure 5.10, left plot: blue and purple datapoints). Chromatin modifier gene mutations (e.g. ASXL1, EZH2), which were seen as the first driver mutation in two controls, but no pre-AMLs, were both acquired under the age of 20 (EZH2 age 0-17, ASXL1 age 0-15) (Figure 5.10, left plot: orange datapoints). Spliceosome gene mutations (e.g. SRSF2, SF3B1) which were observed as the first driver mutation in 3 pre-AMLs, but no controls, were acquired across a range of ages (Figure 5.10, left plot: pink datapoints).

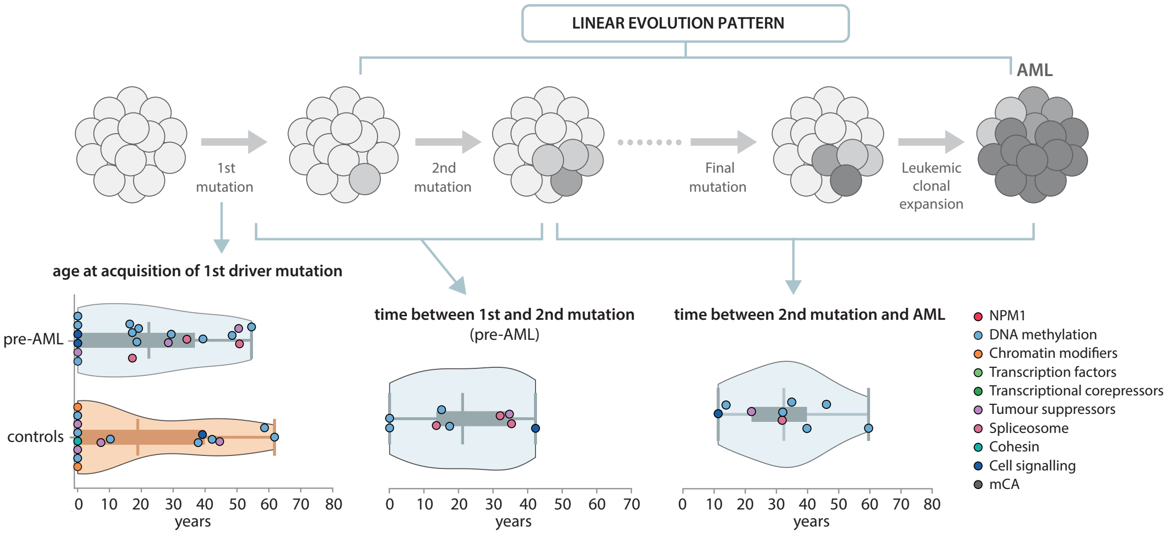


Figure 5.10 Timings for the key steps in the evolution to AML: Linear evolution pattern. The colour of the datapoint indicates the gene class of the mutation. For the time between the first and second mutation, the colour of the datapoint represents the gene class of the first mutation. For the time between the second mutation and AML, the colour of the datapoint represents the gene class of the second mutation. The vertical line in the violin plots represents the mean of the distribution and the thick horizontal bar represents the interquartile range.

Linear evolution: Fitness effects of driver mutations

For samples that showed a linear evolution pattern, there was a range of fitness effects in both pre-AML cases and controls. The average fitness effect of the first driver mutation was higher in pre-AML cases at 22% per year (range 8 - 83%) compared to 16% per year (range 5 - 41%) in controls, although the differences in the distribution of fitness effects did not meet statistical significance with our small set of samples ($p = 0.15$, two sample Kolmogorov-Smirnov test) (Figure 5.11, left plot). After acquisition of the second mutation in pre-AML cases, there was a broad range of double-mutant clone fitness effects. The average fitness effect of the double-mutant clone was 66% per year, but ranged from 15% to 292% per year (Figure 5.11, middle plot). Some of the double-mutant clone fitness effects were greater than would be expected if fitness effects were simply additive and suggests that some mutations behave synergistically to increase fitness.

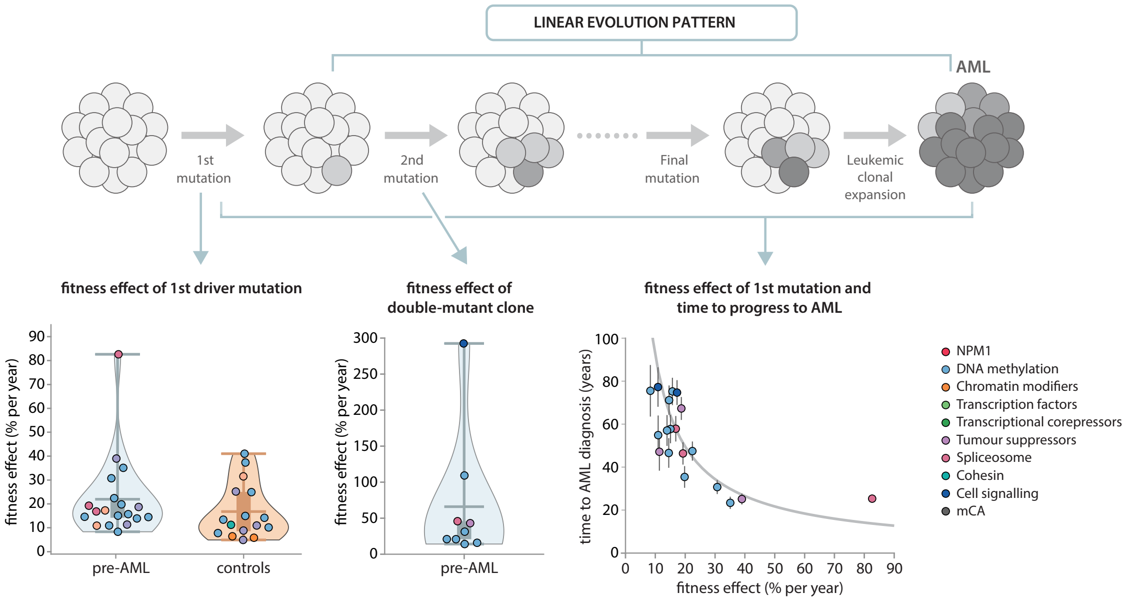


Figure 5.11 Fitness effects of single- and double-mutant clones: Linear evolution pattern. The colour of the datapoint indicates the gene class of the mutation. The vertical line in the violin plot represents the mean of the distribution and the thick horizontal bar represents the interquartile range. The error bars on the time to AML diagnosis are calculated as $\pm 1/s$, where s is the fitness effect per year. The grey solid line on the time to AML diagnosis plot is the expected time to sweep (t_{sw}), based on a mutation's fitness effect, under a simple branching process model of HSC dynamics: $t_{\text{sw}} = \log(N\tau s)/s$.

There was correlation between the fitness effect of the initial driver mutation and time to AML diagnosis, with higher fitness effects associated with shorter times to progress (Figure 5.11, right plot). The time to AML diagnosis was broadly consistent with the expected time for a mutation, with fitness effect s, to sweep (t_{sw}), under a simple branching process model of HSC dynamics (Figure 5.11, right plot, grey solid line). In this model (Chapter 2 Section 2.2, Appendix 1 Section A.1), where beneficial mutations increase the average offspring per HSC generation from 1 to $1+s$, the characteristic clone size, after t generations is:

$$n \sim \frac{e^{st}-1}{s} \quad \Rightarrow \quad n \sim \frac{e^{st}}{s} \quad \text{when } st \gg 1 \tag{5.5}$$

When a mutation has swept, $n = N$, which means:

$$N \sim \frac{e^{st}}{s} \quad \Rightarrow \quad t = \frac{\log(Ns)}{s} \tag{5.6}$$

Considering the time to sweep (t_{sw}) in years, rather than HSC generations, this gives:

$$t_{\text{sw}} = \frac{\log(N\tau s)}{s} \tag{5.7}$$

where τ = time in years between successive symmetric differentiation divisions

This is the line shown in Figure 5.11 (fitness effect vs. time to AML diagnosis).

DNA methylation mutations were seen as both the first and second mutations in 4 of the pre-AML cases with linear evolution. These were DNMT3A/DNMT3A mutations in two individuals and TET2/TET2 or DNMT3A/IDH2 in the other two individuals. For all 4 individuals, the effect of the second mutation on the fitness effect of the clone was only modest, with the greatest increase observed in the DNMT3A/IDH2 sample, whose fitness increased from 14% per year in the DNMT3A single-mutant clone to 21% per year in the DNMT3A/IDH2 double mutant clone (Figure 5.12, light blue/ light blue datapoints). A minimal increase in fitness effect was also observed in the control sample which acquired a chromatin modifier gene mutation (ASXL1) first, followed by a DNA methylation gene (TET2) mutation, with the fitness increasing from only 6% to 9% per year (Figure 5.12, orange/ light blue datapoint). For two of the samples (TET2/TET2 and DNMT3A/DNMT3A), although the similar growth trajectory dynamics suggested the two mutations were in the same clone, the combined VAFs were <50% meaning it was difficult to exclude the possibility that the two mutations were in separate clones. If they were, the reason they have similar fitness effects could be simply because the two mutations have a similar effect within the gene.

The most significant incremental increase in fitness effects from single- to double-mutant clone was observed when the first and/or second mutation was either a spliceosome gene mutation (e.g. SRSF2, SF3B1) and/or a cell signalling gene mutation (e.g. JAK2, MPL), with one of the pre-AML cases increasing their SRSF2 P95H single-mutant clones fitness from 83% per year to 293% per year with the acquisition of an MPL Y591D mutation (Figure 5.12, pink/ dark blue datapoint).

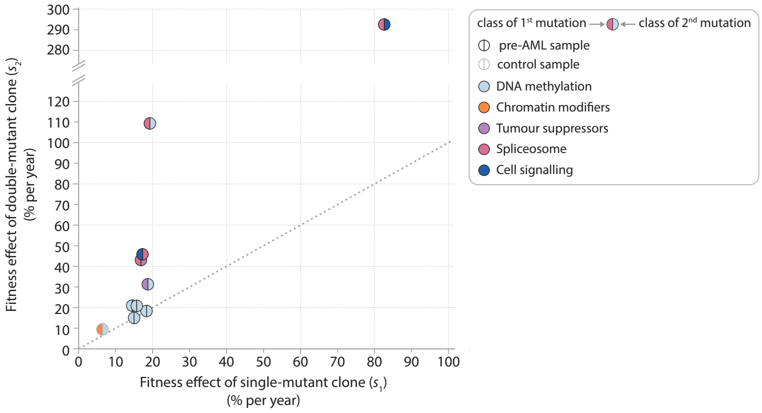


Figure 5.12 Fitness effect of single and double mutant clones in samples with linear evolution pattern. The two colours in each datapoint represent the class of the first (left) and second (right) mutation. Pre-AML samples are represented by black circles and controls are represented by grey circles (only 1 control sample).

5.3.2 Late evolution pattern

Estimating the age at acquisition and fitness effect of the initial driver mutation for samples that show a late evolution pattern was a little more complicated because the mutations were typically only detected at the final timepoint, meaning there was no trajectory to extrapolate. However, we can set an upper bound on the VAF at the preceding timepoint by knowing that it is unlikely to be greater than ∼1/read depth, otherwise it would have been detected (Figure 5.13). Taking this as the VAF at the preceding timepoint enables us to apply the method described for the linear evolution samples (Section 5.3.1), with the understanding that the inferred fitness effects and acquisition ages for these late evolution samples are 'lower bounds' on what they are actually likely to be.

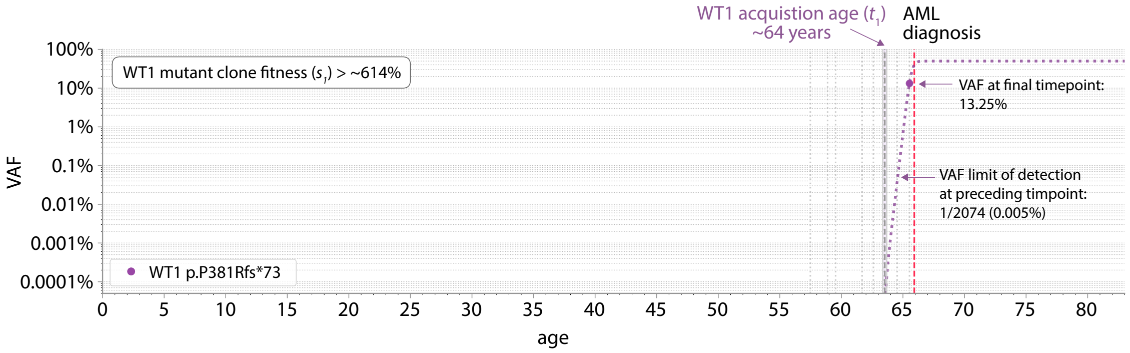


Figure 5.13 Estimation of acquisition age and fitness effects for mutations showing a late evolution pattern. Solid colour datapoint indicates the measured VAF from the blood samples at the final timepoint (each timepoint indicated by a vertical grey dashed line). The VAF limit of detection at the preceding timepoint was used to estimate the maximum VAF at that timepoint. The dashed coloured line indicates the extrapolated trajectory inferred using maximum likelihood approaches, which allow estimation of the acquisition age and fitness effect (s) of the WT1 mutant clone. The error measurement shown for the acquisition ages is $\pm$ 1/s.

Late evolution: Age at acquisition of driver mutations

By applying this method to all the pre-AML samples that showed a late evolution pattern (Appendix D.3), and had a mutation detectable in the final timepoint sample, we find that the average age at acquisition of the first driver mutation was $\sim$ 62 years old (range 46 - 75 years) (Figure 5.10, left plot). We did not observe the acquisition of the second driver mutation in most cases, but we can calculate the average time to progression of AML from the first mutation, which on average is $\sim$ 4 years (range 2 - 6 years) (Figure 5.14, right plot).

Late evolution: Fitness effects of driver mutations

For the pre-AML samples that showed a late evolution pattern, the initiating driver mutations had markedly higher fitness effects than those with a linear evolution pattern, with an average fitness effect of 410% per year (range 99 - 755% per year) (Figure 5.15, left plot). The higher the fitness effect of the first mutation, the shorter the time to diagnosis of AML (Figure 5.15, right plot). As was the case for samples that showed a linear evolution pattern, the time to AML diagnosis was broadly consistent with the expected time for a mutation, with fitness effect s, to sweep (t_{sw}), under a simple branching process model of HSC dynamics (Figure 5.15, right plot, grey solid line).

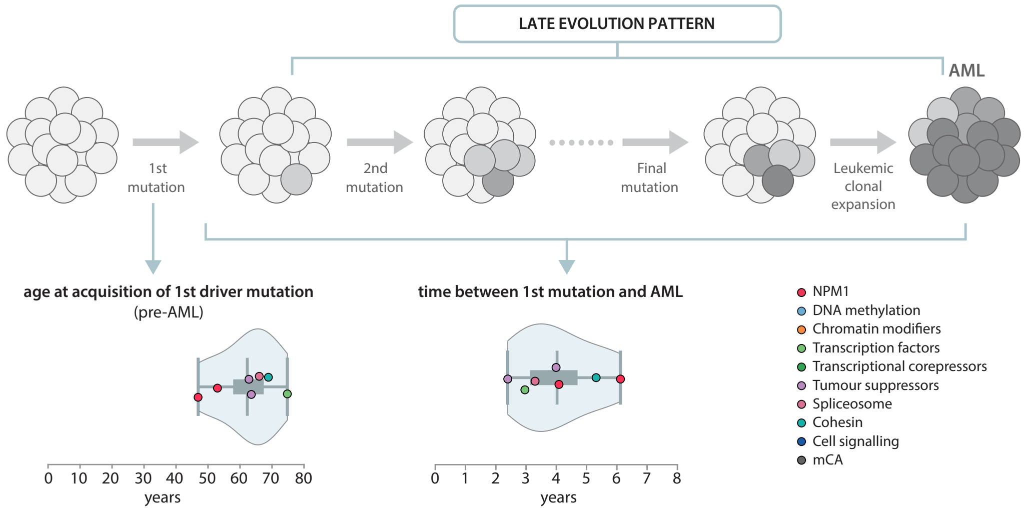


Figure 5.14 Timings for the key steps in the evolution to AML: Late evolution pattern. The colour of the datapoint indicates the gene class of the mutation. The vertical line in the violin plot represents the mean of the distribution and the thick horizontal bar represents the interquartile range.

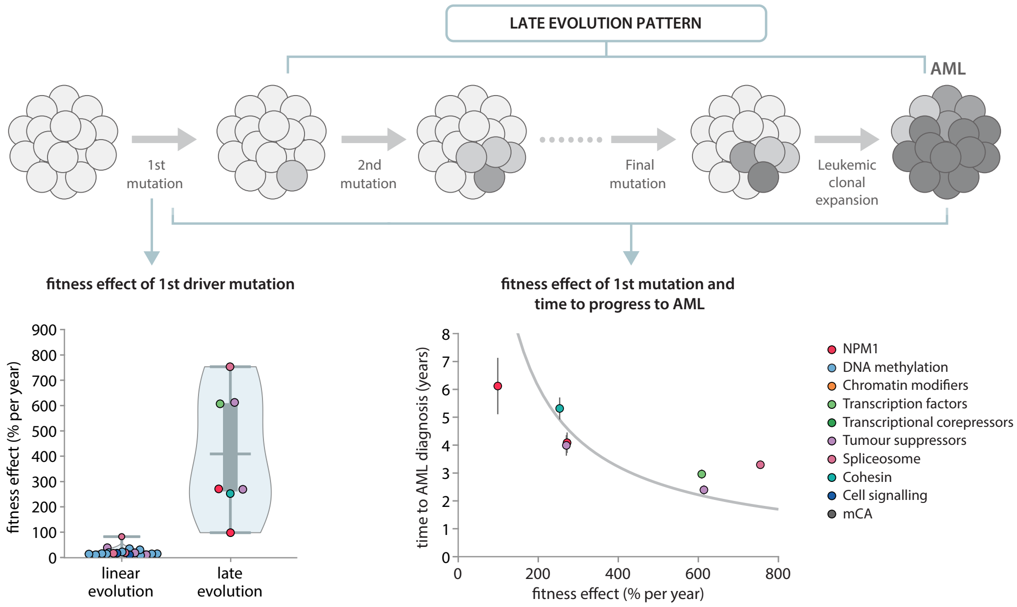
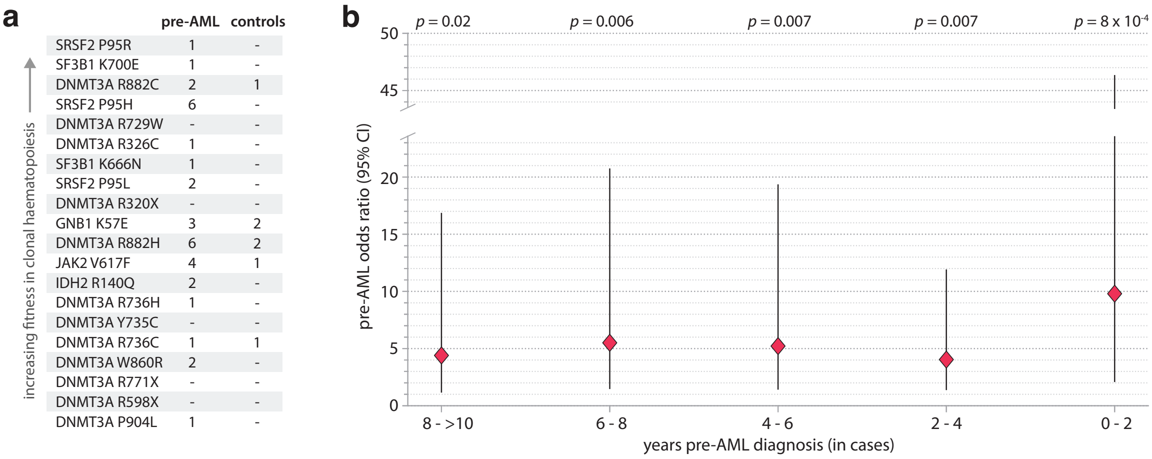


Figure 5.15 Fitness effects of single- and double-mutant clones: Late evolution pattern. The colour of the datapoint indicates the gene class of the mutation. The vertical line in the violin plot represents the mean of the distribution and the thick horizontal bar represents the interquartile range. The error bars on the time to AML diagnosis are calculated as $\pm 1/s$, where s is the fitness effect per year The grey solid line on the time to AML diagnosis plot is the expected time to sweep (t_{sw}), based on a mutation's fitness effect, under a simple branching process model of HSC dynamics: $t_{\mathrm{sw}} = \log(N\tau s)/s$.

5.4 Highly fit clonal haematopoiesis variants are enriched pre-AML

While many of the genes mutated in pre-AML cases were also mutated in controls, the specific variants detected in pre-AML tended to be known ‘hotspot’ sites recurrently mutated in haematological malignancies and/or variants we inferred to be highly fit in clonal haematopoiesis. Indeed, the group of 20 variants we inferred to be highly fit in clonal haematopoiesis (Chapter 2)[93] were significantly enriched in pre-AML cases compared to controls, even 8 to >10 years pre-diagnosis (Figure 5.16). The odds ratio of detecting one or more of these variants was 4.4 (95% confidence interval 1.1 to 16.9) at 8 to >10 years pre-diagnosis, which increased to an odds ratio of 9.8 (95% confidence interval 2.1 to 46.4) in the 2 years pre-diagnosis.



	pre-AML	controls
SRSF2 P95R	1	-
SF3B1 K700E	1	-
DNMT3A R882C	2	1
SRSF2 P95H	6	-
DNMT3A R729W	-	-
DNMT3A R326C	1	-
SF3B1 K666N	1	-
SRSF2 P95L	2	-
DNMT3A R320X	-	-
GNB1 K57E	3	2
DNMT3A R882H	6	2
JAK2 V617F	4	1
IDH2 R140Q	2	-
DNMT3A R736H	1	-
DNMT3A Y735C	-	-
DNMT3A R736C	1	1
DNMT3A W860R	2	-
DNMT3A R771X	-	-
DNMT3A R598X	-	-
DNMT3A P904L	1	-

Figure 5.16 Highly fit clonal haematopoiesis variants are enriched in pre-AML samples. a. Total number of pre-AML and control samples with each of the highly fit clonal haematopoiesis variants (from Chapter **??** and Watson et al 2020[93]). **b.** Odds ratios (with 95% confidence intervals) of detecting one or more of the 20 highly fit variants in pre-AML vs control samples.

5.5 Detection of AML-specific mutations pre-AML diagnosis

Whilst many pre-leukaemic mutations are also found in clonal haematopoiesis, there are some mutations, such as FLT3-ITD and NPM1 exon 12 frameshift mutations, that have only ever been found in individuals with AML[20,142,143]. The absence of these mutations in healthy individuals, together with evidence from single cell sequencing studies, suggests that these mutations occur ‘late’ in AML development, but how late they occur is unknown. Using our deep error-corrected sequencing approach we were able to detect NPM1 exon 12 frameshift mutations in 3 individuals (1 with linear evolution pattern, 2 with a late evolution pattern) and using our method to infer acquisition age and fitness effects (Section 5.3) could infer that these mutations occurred $\sim$ 4 - 6 years pre-diagnosis. We detected a FLT3-ITD mutation in 1 individual (with a late evolution pattern), but as discussed in Chapter 4, there may also be additional individuals with FLT3-ITD mutations (RT-PCR validation of these samples is still pending).

5.6 Discussion

Having access to blood samples collected at multiple timepoints pre-AML diagnosis, combined with an error-corrected sequencing approach that allows low frequency detection of a comprehensive array of genetic changes, provides us with the ability to study the pre-leukaemic evolutionary process over time. This allows us to see the progression to AML unfold, from acquisition of the initial driver mutation through to just before diagnosis of AML. Recent single cell sequencing studies have inferred AML evolutionary history using samples collected at the time of diagnosis[206,207], and whilst this enabled reconstruction of pre-AML lineages, it did not provide any information on mutation timings. Also, because AML diagnosis samples are naturally dominated by the AML clone, the presence of intermediate clones that were outcompeted during the course of preleukaemic evolution may be missed, resulting in an incomplete picture of the preleukaemic evolutionary process.

Four different patterns of pre-leukaemic evolution can all lead to AML

By analysing the clonal trajectories generated from blood samples collected at multiple timepoints pre-AML, we found four main patterns of pre-leukaemic clonal evolution, which could all give rise to AML: linear evolution, evolution with clonal competition, late evolution and static evolution. The 'linear evolution' pattern was most common and is consistent with the stereotypical model of cancer evolution[2,208], where sequential acquisition of driver mutations occurs, with each new driver mutation increasing the fitness effect of the clone, which causes it to outcompete the preceding clones. In this evolution pattern the limiting factor in progression appears to be the time it takes for the next mutation to occur. In contrast, in the 'clonal interference' pattern of evolution there is no shortage of mutations and the limiting factor in progression appears to be waiting for clones to outcompete each other to become the dominant clone. Clonal interference is a common evolution pattern observed in a number of cancer types and is typically associated with poorer survival[209] and increased therapy resistance[210]. The mechanisms underlying the clonal interference process remain unclear, although some have considered clonal interference to be a proxy for greater mutational instability, faster evolution and/ or interclonal co-operativity[211]. Clonal interference is expected to occur when $N\tau\mu \gg 1$[45,212], which suggests the mutation rate (μ) and/ or HSC population size ($N\tau$) may be increased in the individuals whose pre-AML evolution showed a clonal interference pattern.

In the 'late evolution' pattern, pre-AML samples appear genetically indistinguishable from control samples, up until 1-2 years prior to AML diagnosis, when the emergence of highly fit driver mutations ('uber drivers') results in significant clonal expansion. The limiting factor in progression in 'late evolution' appears to be the time it takes for this driver mutation to occur. This 'late evolution' pattern is similar to the 'punctuated' evolution pattern described in other cancer types, such as breast[213] and prostate cancer[214], which are thought to be caused by a sudden catastrophic genomic event such as chromothripsis[215,216], chromoplexy[214] or kataegis[57], which occur on a permissive anti-apoptotic background. 'Punctuated evolution' is typically characterised by chromosomal structural rearrangements, however, which we did not observe in any of the pre-AML samples with late evolution. This suggests that, in AML, stochastic acquisition of a highly fit mutation may be sufficient to trigger the initiation of rapid evolution to AML. It is possible that we are missing key events, which occur earlier in the pre-leukaemic evolution in these individuals, due to our targeted panel. Whole-genome sequencing of the samples from these individuals may be helpful to determine whether this is the case.

In the 'static evolution' pattern clones show no growth at all, and yet must have had at least a moderate fitness effect to have reached the VAFs at which they were observed. This finding has also been observed in other pre-AML studies[77] and a slowing of DNMT3A clonal expansion with age has recently been reported in clonal haematopoiesis[103]. Whilst this could be due to an unseen clonal competitor not covered by our panel, the competitor would need to have the same fitness effect as the static clone in order to explain the unchanging VAF measurements. Whilst this is not implausible, it seems unlikely that this is the explanation in all of the samples that showed this pattern of evolution.

Another possible explanation for the static evolution pattern is that the mutated cell of origin is not an HSC, but a long-lived lineage-restricted progenitor cell (eg. common myeloid progenitor) that also has the ability to self-renew. This would cause the mutations to be lineage-restricted and so their clonal expansion could be confined to the proportion of blood cells derived from that lineage. Whilst single-cell sequencing could ascertain if this was the case, this is unfortunately not possible with our samples. Instead, DNA methylation patterns could be analysed, to attempt to estimate the proportion of different haematopoietic cell types present[217] and determine if this was consistent with the cell fraction in which the clonal expansion appeared to be constrained.

Another explanation is that an external influence, such as the immune system, 'holds back' clonal expansion once it reaches a certain size, until a final AML-triggering mutation occurs. This 'holding back' of clonal expansion is analogous to 'tumour mass dormancy', which has been described in other cancers[218,219] and can be mediated by the immune system or inefficient nutrient supply to the cells. In this scenario, clonal growth is still occurring, but at the same rate as immune (or hypoxia)-mediated cell death. Whilst this is possible, clonal growth and cell death would need to be perfectly balanced, to explain the static trajectories, which seems perhaps unlikely. 'Cellular dormancy' is also a possibility, in which cells transition to a quiescent, cell cycle-arrest state, but retain the capacity for self-renewal when reactivated. This has been described in chronic myeloid leukaemia (CML) leukaemia stem cells, mediated by 'supportive signals' secreted by non-transformed bone marrow niche cells[220–222]. The bone marrow microenvironment/ niche provides regulatory signals that tightly control self-renewal, quiescence, differentiation and migration of haematopoietic stem cells[223] and so may play an important role in promoting 'cellular dormancy', and thus a 'static' pattern of evolution, in some individuals.

Whether periods of stasis occur in all pre-leukaemic clones, but only coincided with the timings of some of the samples, or whether specific genetic or epigenetic factors predispose to stasis is unclear and would be an interesting area for future research. Understanding this better might also shed light on why clonal competition seemed to occur within a constrained fraction of cells in some samples that showed a ‘clonal interference’ pattern of evolution.

Highly fit initiating driver mutations result in faster progression to AML

One of the principles underlying our understanding of pre-cancerous mutation acquisition and clonal expansion is that the greater the fitness effect of a mutation, the faster the clone will expand and the more likely it is that subsequent mutations will be acquired within the same clone. Consistent with this, we found that the higher the fitness effect of the initial driver mutation, the shorter the time to progress to AML. This was particularly evident for samples that showed a late evolution pattern, in which the initial driver mutations had fitness effects as high as 600-800% per year. Fitness effects this great would result in the clone size doubling approximately every 6 weeks.

Interestingly, for both ‘linear’ and ‘late’ evolution samples, the relationship between the fitness effect of the first mutation and the time to progress to AML was broadly consistent with the predicted amount of time it would take for the mutation to sweep, under a simple branching process model of HSC dynamics. This suggests that once the initial mutation has swept, progression to AML occurs shortly after. Given we can see a gradual step-wise accumulation of mutations in the linear evolution pattern, and the average time from acquisition of the 2nd mutation to AML diagnosis is $\sim$ 30 years, this finding is hard to explain and further work is required to explore this further.

Early detection of AML-specific mutations

Using deep error-corrected duplex sequencing, we were able to detect NPM1 and FLT3 mutations up to 2 years pre-AML diagnosis. Whilst some of these cases were individuals that showed ‘late’ evolution, who would likely be unidentifiable as high-risk until these mutations occurred, some of these mutations occurred in individuals that had already acquired several mutations, which had expanded to high VAF. Detection of NPM1 or FLT3 mutations 1-2 years before diagnosis in these individuals potentially provides a window of opportunity for early intervention studies, to try to prevent progression to overt AML. Our early detection of these mutations also highlights the benefit afforded by deep error-corrected low VAF variant calling, particularly in high-risk individuals.

Initial driver mutations can occur at any age

Using multi-timepoint clonal trajectories and insights from evolutionary theory, we were able to estimate the age at acquisition of the first AML driver mutation. In individuals with a linear evolution pattern, the average age at acquisition was $\sim$ 22 years old, but could occur anywhere between 0 and 62 years old. In individuals with a late evolution pattern, the average age at acquisition was $\sim$ 62 years old, which was, on average, only $\sim$ 4 years before AML diagnosis. These results highlight that, whilst the presence of multiple pre-leukaemic mutations is certainly associated with an increased risk of AML, the absence of detectable pre-leukaemic mutations at, e.g. age 60, does not necessarily mean that an individual is not at risk. It also suggests that early detection may not be possible in all cases of pre-AML.

Age- and sex- differences in AML development

AML in older individuals ($\geq$60 years old) has been found to be genetically distinct from AML in younger individuals[224,225], with higher rates of mutations in spliceosome components, epigenetic regulators and in DNA repair factors[226]. The average age at AML diagnosis in our UKCTOCS cohort was 71 years old and so it is important to bear in mind that our findings may not be generalisable to younger patients with AML. Similarly, sex biases are known to occur for certain pre-leukaemic gene mutations and AML is significantly more common in men than women[201]. In particular, DNMT3A, TP53, NPM1 and FLT3-ITD mutations are known to be more common in women[77,227] and ASXL1, SRSF2 and U2AF1 more common in men[227]. The mechanism underlying these gender biases is not fully understood, but may be related to bone marrow microenvironment differences between men and women, as suggested by the different mutational signatures observed between men and women for certain pre-leukaemic gene mutations[227]. All of the participants of the UKCTOCS are women and so it is important to bear this in mind if attempting to generalise these findings to men as well.

Future work

Analysis of the data generated from this unique set of pre-leukaemic longitudinal samples is still ongoing. Thus far we have inferred clonal composition by simply using VAF measurements and clonal trajectory dynamics, a benefit afforded by having longitudinal samples. For example, observing two or more mutations whose combined VAFs were >50% (in the absence of copy number change) at later timepoints allowed us to infer that these mutations were present in the same clone across all preceding timepoints. For mutations whose combined VAFs were not >50% at any timepoint, we could infer clonal co-localisation if their growth trajectories were the same, or indeed clonal independence if they showed divergent trajectories. Whilst this approach allowed us to infer the clonal composition for many of the samples, there were some mutations, e.g. those present at low VAFs across all timepoints, for which this was not possible. Bayesian clustering approaches (e.g. Dirichlet process models)[228–230], directed acyclic graph networks[231,232] and matrix deconvolution frameworks[233] are all possible methods that can be used to estimate phylogenetic relationships between clones and determining whether adaptations of these approaches could be applied to our longitudinal samples is an area for future work.

At the moment we have only determined the acquisition age and fitness effects for the first two mutations and another area for future work is to extend this to include additional subsequent mutations as well.

AML can be defined as ‘secondary’ (s-AML), ‘therapy-related’ (t-AML) or ‘do novo’. s-AML arises following transformation of antecedent myelodysplastic syndrome (MDS) or myeloproliferative neoplasm (MPN), t-AML arises after exposure to leukaemogenic therapy and de novo AML arises without any known exposure or preceding condition. It is possible that some of the UKCTOCS participants developed MDS or MPN before they developed AML. Indeed, 20% of our pre-AML cases had suspected biallelic TET2 mutations, which are often associated with chronic myelomonocytic leukaemia (CMML)[203] and 42% of the pre-AML cases had mutations in SRSF2, SF3B1, U2AF1, ZRSR2, ASXL1 or EZH2, which have been previously shown to be >95% specific for a diagnosis of secondary AML[234]. It is also possible that some of the participants may have received leukaemogenic therapy for a previous solid cancer. Unfortunately we will not be able to obtain information on blood counts/ haematological parameters at any of the timepoints, but we can obtain information from the Office of National Statistics (ONS) National Cancer Registry and NHS Hospital Episode Statistics (HES) database regarding any antecedent diagnoses and should also be able to identify if there had been any prior leukaemogenic therapy. This will be important for determining whether particular pre-leukaemic evolution patterns are associated with particular types of AML.

Our ratio of pre-AML cases to controls is currently 1:1. Given the rarity of AML (4.3 cases per 100,000 individuals per year[201] relative to clonal haematopoiesis, and therefore the low risk of progression of clonal haematopoiesis to AML, future work involving the analysis of a greater number of controls should be considered. The associated increased cost that this would incur would also need to be taken in to account.

At the moment, the closest samples we have to AML diagnosis are $\sim$ 3 months prior, which means we are unlikely to have visualised the full evolutionary history preceding AML. We are attempting to obtain surplus bone marrow aspirate slides from the time of AML diagnosis, which would be invaluable for determining the events that trigger the final stage of evolution to AML.

6

Summary & Discussion

6.1 Summary and Discussion

Whilst clonal haematopoiesis has been described as an inevitable consequence of ageing[9], over the past decade its potential clinical significance as a precancerous state has become increasingly recognised, representing the first of a multistep process that in some individuals can progress to a blood cancer, such as AML[3,4,29]. The hope is that if we can identify high risk individuals, we might be able to prevent progression to AML, which, once it occurs, sadly claims the lives of 70-80% of patients within five years of diagnosis[201].

To identify the individuals most at risk of progression to AML, it is important to understand the evolutionary dynamics of clonal haematopoiesis in the years, or decades, before AML occurs and how this differs from the dynamics of clonal haematopoiesis in individuals that remain cancer-free. To do this we set out to gain a quantitative understanding of each stage of the step-wise process to AML, from acquisition and clonal expansion of the initial driver mutation (using data from ∼50,000 - 500,000 individuals) all the way through to pre-leukaemic evolution (using longitudinal blood samples collected in the decade preceding AML diagnosis).

6.1.1 Acquisition and expansion of the initial clonal haematopoiesis mutation

Clonal haematopoiesis can result from clonal expansion of single nucleotide variants, insertions/ deletions (indels) or larger chromosomal changes such as mCAs[3,4,31,35]. Over the past decade, the decreasing costs of high-throughput sequencing have resulted in a dramatic increase in the amount of blood sequencing data generated from a number of large population-based studies, such that we now have single nucleotide variant data from the blood of >50,000 individuals[3,4,6–9,38,39,75] and mCA data from ∼500,000 individuals[35]. Having clonal haematopoiesis data from this many people provided us with the unparalleled opportunity to study how mutation rates, genetic drift and fitness differences combine to shape the genetic diversity observed in healthy blood and provided insights in to the first stage of the multi-step process that in some individuals can lead to AML.

Using evolutionary models to understand haematopoietic stem cell dynamics

Models for measuring and predicting clonal evolution are often adapted from the field of evolutionary theory and population genetics, where the principles of mutation, fitness and genetic drift are equally as important[43–46]. These models often build on evolutionary principles proposed by Darwin and Mendel[235] and enable us to not only test underlying model assumptions, but also infer key parameters, such as mutation rates, fitness effects and population size (HSC numbers).

In the work presented here, we considered a simple branching process model of HSC dynamics, based on evolutionary theory[43–46], but adapted to take into account a spectrum of ages and a spectrum of fitness effects (Chapter 2). In this model, HSCs make the stochastic decision to divide either symmetrically (self-renewal or differentiation) or asymmetrically with each cell division. Mutations are acquired at a constant rate throughout life and each mutation confers a fitness effect, which, if > 0 (non-neutral), biases the HSC cell fate towards symmetric self-renewal. This eventually results in exponential growth of the mutant cells, at a rate determined by the fitness effect. This simple model produces predictions for the clone size (VAF or cell fraction) distribution across individuals, as well as how the population prevalence of mutations should increase with age.

We found that the VAF spectra of single nucleotide variants (from ∼50,000 individuals) and the cell fraction spectra for mCAs (from ∼500,000 individuals) were consistent with our simple model of HSC dynamics (Chapter 2 and Chapter 3), revealing that positive selection, not drift, is the major force shaping clonal haematopoiesis. Mutations appear to be acquired at a roughly constant rate throughout life and, if they confer a positive fitness, result in exponential growth. Contrary to the view that clonal haematopoiesis is driven by ageing-related alterations in the stem cell niche[202], the data are consistent with the age dependence of clonal haematopoiesis being driven simply by a continuing risk of mutations and subsequent clonal expansions that lead to increased detectability at older ages. These findings have subsequently been corroborated using single-cell-derived colonies of HSC/MPPs from individuals across a range of ages[92], as well as longitudinal clonal haematopoiesis data[103].

HSC numbers and division rates

Having an estimate for the total number of HSCs is important for developing quantitative models of cancer risk, but this number has historically been hard to measure, with most attempts involving extrapolation from animal models or calculation of relative proportions of cells, with estimates ranging from 10^2 to 10^9 cells[60–67]. Our framework allows us to measure this. We estimate $N\tau$ (HSC number × years between symmetric differentiation divisions) to be $100,000 \pm 30,000$, which is in close agreement with other recent estimates from single cell phylogenies[48,92]. Estimating N separately from τ is challenging (for all population genetic analyses), but by estimating bounds for τ from mutation rates in development and adulthood[48], we can estimate that the total number of HSCs is between 25,000 and 1.3 million (Chapter 2).

Identification of high risk clonal haematopoiesis mutations

Determining mutation growth rates usually requires multi-timepoint data, but we showed that it is possible to quantify the fitness effects of individual variants and mCAs from single timepoint data using our framework. We present a league table of the fittest and potentially most pathogenic variants and mCAs and quantify the distribution of fitness across key clonal haematopoiesis genes. Many of the fitness effects we inferred are consistent with the fitness effects subsequently inferred, by us (Chapter 5) and others[103], using longitudinal blood samples. We found mCA loss events and spliceosome variants (SF3B1 and SRSF2) to be some of the fittest mutations, with fitness effects as high as ∼23% per year, which would result in a doubling of the number of effected stem cells every ∼ 3.5 years. We were able to infer the fitness effect of ∼ 60% of all possible mCAs, of which most were highly fit with fitness effects in the range 10-20% per year. In contrast, we found the distribution of fitness effects for variants across DNMT3A, TET2, ASXL1 and TET2 to be highly skewed, with most mutations in these genes conferring either a weak or no fitness effect. The 20 highly fit variants that we identified conferred a significantly increased risk of AML and we found correlation between mCA fitness and risk of subsequent blood cancer (Chapter 2 and Chapter 3).

The mechanism by which clonal haematopoiesis variants lead to clonal expansions is not fully understood. Cancer-associated mutations are classically categorised as ‘tumour suppressor’ genes or ‘oncogenes’. Whilst JAK2, GNAS and GNB1 could be considered oncogenes, and TP53 a tumour suppressor gene, mutations in these genes account for <10% of the mutations observed in clonal haematopoiesis[3,4,6–9,38,39,75]. The vast majority of clonal haematopoiesis mutations affect genes involved in DNA methylation, epigenetic regulation or RNA splicing. One theory proposes that mutations in these genes result in altered expression of genes involved in HSC self-renewal[236]. Another theory proposes that by altering DNA methylation and chromosomal architecture of enhancer regions, mutations in these genes disrupt differentiation and lineage specification[236]. Elucidating the mechanisms by which these genes cause clonal expansions will be important for designing potential therapeutic strategies.

Deviation from model expectations reveals sex-specific differences in behaviour

Having ascertained that the blood sequencing data, across ∼50,000-500,000 individuals, was consistent with a simple model of HSC dynamics, our framework served as a useful null model to identify variants or mCAs whose behaviour deviated from model predictions. In contrast to the strong age dependence observed in single nucleotide variant prevalence in blood, we found the age dependence of some mCAs was more variable, particularly in women. This pointed to a violation of the underlying assumptions of our model and suggests that the risk of acquisition and/or expansion of certain mCAs is non-uniform throughout life and is influenced by gender-specific factors (Chapter 3).

Other evidence already exists for gender-specific factors playing a role in mutation acquisition and/or clonal expansion in clonal haematopoiesis and AML. AML, for example, is significantly more common in men and this male predominance increases further with age[201]. Certain genes, such as ASXL1, BCOR, RUNX1, SRSF2 and U2AF1 are more commonly mutated in men, whereas others such as TP53, DNMT3A, NPM1, and FLT3-ITD are more commonly mutated in women[227]. The precise mechanism behind these gender-differences and whether they are caused by mutation rate or fitness differences is not known. Higher levels of bone marrow fat in younger men compared to younger women has been proposed as a theory[227,237]. Another theory, based on the different mutational signatures observed between men and women in certain genes, is that tissue inflammation, or activity of AID/APOBEC may play a role[227]. Further work is needed to ascertain the cause of these gender differences and will be important for building quantitative models of cancer risk as well as potential therapeutic interventions.

Model limitations

It is important to bear in mind that our model, as with any model, is a simplified version of what is likely a complex, intricately controlled process. Whilst our model appears to capture the behaviour of mutations across all ∼50,000-500,000 individuals it involves a number of assumptions that may not be valid at an individual level in certain situations.

One assumption our model makes is that each mutation has a specific mutation rate and that this rate is constant throughout life. Studies involving solid tumours have shown that this may not always be the case and has provided evidence that mutational processes can cluster in time[238–240].

Another assumption our model makes is that beneficial mutations (with a fitness effect >0) cause cells to grow exponentially. Whilst we know from our longitudinal pre-AML and control samples that this seems to be true most of the time, we do see clear evidence of situations when this is not the case (e.g. pre-AML and control samples that showed a static evolution pattern, Chapter 5). We have not identified a particular gene or mutation that behaves in this way and so it is possible that cell-extrinsic factors are the cause. Other evidence for non-exponential growth comes from a recent study showing that the growth of certain DNMT3A mutant clones slows down with age, in the context of an increasingly competitive oligoclonal landscape[103].

Our model also does not take in to account the effect of ageing on HSC behaviour or fitness effects. Ageing has been shown to have numerous effects within the haematopoietic system, including myeloid proliferation bias[241], decreased bone marrow cellularity[242], reduced lymphopoiesis[243] and reduced erythropoiesis[244]. Previous work has also shown that HSC self-renewal and differentiation capacity reduce with age[245,246], such that any ageing HSCs that have retained their ability to self-renew are most likely to expand. Ageing has also been shown to affect fitness, with B-lymphopoiesis showing a dramatic decrease in fitness with age[247]. The increasing acquisition of mutations with age is also thought to impact the fitness of mutations[248], most likely due to competition between clones. Whilst the clonal haematopoiesis data across ∼50,000-500,000 individuals appears to be consistent with a simple model in which the fitness effect conferred by a mutation is constant throughout life, more complex models that allow for age-related variations in fitness effects could be used to further validate this. These more complex models would require a greater number of parameters, however, and it is likely that data from more individuals, and from multiple timepoints, would be required to allow robust discrimination between models.

Another limitation of our model is that it focuses on cell-intrinsic fitness effects. While the data across ∼ 50,000-500,000 individuals is quantitatively consistent with cell-intrinsic fitness effects playing the major role in shaping the variation in VAFs and cell fractions that we see between individuals, it is important to bear in mind that cell-extrinsic effects such as chemotherapy[75,96–98], acute infection[99,100] and inflammation[101] likely play a role in certain contexts. Indeed, variants in certain genes (e.g. PPM1D, TP53, CHEK2 and ASXL1) have been shown to be strongly influenced by external factors[97,98,102].

There are also features of the data that our simple model cannot explain. One of these is the larger than expected number of mutations observed between individuals (ranging from 0 to >10 for the same detection limit)[7], which has also been observed in other tissues. Indeed, recent work involving the bladder epithelium, has shown extensive inter-person variation in mutation burden, mutational signatures and selection, which may be due to variability in environmental exposures or genetic background, or an interplay between the two[249]. Differences in germline predisposition may also be important. Indeed, germline loss of MBD4 has been found to predispose to leukaemia due to a mutagenic cascade driven by 5-methylcytosine (5-mc)[250]. These individuals sustain high levels of damage from 5mc deamination and experience clonal expansions decades earlier that age-matched counterparts with normal MBD4.

6.1.2 Progression of clonal haematopoiesis to AML

Analysing blood samples across thousands of individuals, unselected for haematological malignancy, provided us with insights into the first stage of the multistep process that precedes AML development. To determine how and why some individuals then acquire additional mutations and progress to AML, we identified longitudinal blood samples that had been collected annually at multiple timepoints (mean 5, range 2-11) from 50 individuals who subsequently developed AML, as well as 50 age-matched controls who remained blood cancer free. We developed a custom comprehensive targeted NGS panel, which could detect an array of clonal haematopoiesis and AML-associated genetic changes, including gene mutations, mCAs and chromosomal rearrangements. To ensure we could detect mutations as far back in time as possible, we used duplex error-corrected sequencing and developed a custom *in silico* noise correction method, which allowed us to call variants down to single molecule resolution. We developed a custom chromosomal rearrangement caller, for accurate translocation and inversion VAF estimation and harnessed the power of the longitudinal samples to phase SNPs on an individual basis, enabling us to call mCAs at cell fractions as low as 0.1% (Chapter 4).

AML as an evolutionary process

By reconstructing the clonal evolution trajectories for all 50 pre-AML and control samples we found there were four main pre-leukaemic evolution patterns (Chapter 5). ‘Linear evolution’ was most common and involved the successive acquisition of mutations, with each new mutation increasing the fitness of the clone, causing it to outcompete the preceding clones. This pattern of evolution is consistent with the stereotypical model of cancer evolution[2,208] and is typical of small-to-moderate-sized HSC populations in which beneficial mutations are sufficiently rare[45]. In the ‘clonal interference’ pattern, multiple clones were present with clear evidence of clonal competition between them, in a pattern of evolution that is typical of large HSC populations and/or high mutation rates[45]. In the 'late evolution' pattern, pre-AML samples were indistinguishable from controls up until 1-2 years before AML diagnosis, when the emergence of highly fit driver mutations resulted in considerable clonal expansion. This pattern of evolution is similar to a punctuated evolution pattern observed in other cancer types, although this is usually caused by a sudden catastrophic genomic event. In the static evolution pattern, clones that had already expanded to a moderate size appeared to have stopped growing and showed no growth at all across the sample timepoints.

Static evolution has also been observed in other studies[3,251] and several theories exist to explain why this occurs. Most theories focus on either cell-extrinsic factors[218,219], such as the immune system or inefficient nutrient supply, or cell-intrinsic factors, in which cells transition to a quiescent, cell-cycle arrest state[220–222]. The bone marrow microenvironment provides regulatory signals that can trigger quiescence and so may play an important role[223]. Indeed, there is some evidence that the microenvironment can help to stop cancer developing[252].

Constrained clonal evolution

Another, perhaps more plausible explanation for some cases of static evolution, is that the mutations have occurred within a long-lived lineage-restricted progenitor cell with the ability to self-renew. This may also be the explanation for some of samples that showed a clonal interference evolution pattern, but in which the clonal competition appeared to be constrained to within 40-60% of the cells.

Whilst HSCs are classically considered the cell of origin for clonal haematopoiesis, there is some evidence that this may not always be the case. Clonal haematopoiesis mutations are often found in granulocytes, monocytes and NK cells, but are not always present in B cells and are rarely present in T-cells[253]. There may also be a gene-specific pattern to the cell type in which mutations are found. For example, JAK2 mutations are found in the T-cells of the majority of MPN patients[254–256], whereas DNMT3A mutations are found in the T-cells of only 30-50% of individuals with clonal haematopoiesis[253,257] and TET2, ASXL1 and SF3B1 are rarely found in T-cells[253]. Whilst these findings could provide evidence of mutations occurring in myeloid-lineage restricted progenitors, it is also possible that the mutations simply cause a myeloid bias[258]. It is also possible that the mutations which are not seen in T-cells are mutations that arose in middle-age, after the age at which T-lymphopoiesis has largely ceased[258]. Against this is the finding that, in some of our longitudinal samples that showed a linear evolution pattern, TET2 mutations appeared to be acquired within the first decade of life (i.e. before T-lymphopoiesis has ceased).

Fitness effects correlate with time to progress to AML

We find that the fitness effect of the initial driver mutation correlates with the time it takes to progress to AML. This was particularly evident for mutations in the 'late evolution' samples in which fitness effects were as high as 600-800% per year. Interestingly, for both 'linear' and 'late' evolution samples, the relationship between the fitness effect of the first mutation and the time to progress to AML was broadly consistent with the predicted amount of time it would take for the mutation to sweep, under a simple branching process model of HSC dynamics. This suggests that once the initial mutation has swept, progression to AML occurs shortly after. Given we can see a gradual step-wise accumulation of mutations in the linear evolution pattern, and the average time from acquisition of the 2nd mutation to AML diagnosis is $\sim$ 30 years, this finding is hard to explain (Chapter 5). Further work is required to explore this further.

Early detection of AML

The 'holy grail' of clonal haematopoiesis is the ability to definitively predict who will progress to AML. This would allow us to identify high risk individuals who need closer monitoring and may benefit from early intervention, whilst safely reassuring low risk individuals. Because clonal haematopoiesis is so common[7,93] and AML is relatively rare (4.3 cases per 100,000 individuals per year[201]) it is important that any early detection test has a very high positive predictive value, otherwise the false positive rate could be considerable.

Using our longitudinal samples we found that the initial driver mutation could occur at any age. For example in the linear evolution pattern, the average age of mutation acquisition was $\sim$ 22 years old, but could occur anywhere between 0 and 62 years old. In individuals with a late evolution pattern, the average age at acquisition was only $\sim$ 4 years before AML diagnosis. Given these individuals were indistinguishable from controls >2 years pre-AML diagnosis, it seems that early detection of AML may not be possible in all individuals. Indeed, in Jaiswal et al's population-based study, several individuals with longitudinal samples that were diagnosed with a haematological malignancy, within an 8 year period of follow-up, did not have any evidence of clonal haematopoiesis in their baseline sample[3], although some of this may be attributable to them predominantly focusing on hotspot variants and their VAF limit of detection, which was only 3.5%.

For those that do have clonal haematopoiesis mutations detectable, a key challenge is determining who is high risk. Recent studies have shown that particular gene mutations are associated with increased risk[76,77] and we can see from our analyses that the presence of one or more of our 'top 20 fittest variants' is associated with particularly high risk. These 'top 20' variants are significantly enriched in pre-AML individuals, even 8-10 years pre-diagnosis (Chapter 2 and Chapter 5). Clonal complexity is also associated with increased risk[76,77], although this may only be significantly different between cases and controls <8 years pre-AML diagnosis (Chapter 5). Higher VAF mutations have been consistently associated with greater risk[76,77], but whilst studies have not found an association between very low VAF mutations and AML, this could be because the study follow-up length was insufficient. Depending on the age at acquisition of the mutation, however, the individual may pass away from another cause before they found themselves at imminent risk of AML. Using ultra-deep error corrected sequencing of longitudinal samples we found that it is possible to detect NPM1 mutations, which characteristically occur late in AML development and have never been seen in individuals who do not progress to AML, as early as 2 years pre-AML diagnosis. Whilst ultra-deep error-corrected sequencing will not be cost-effective in all individuals, this finding highlights the benefit afforded by low VAF variant calling, particularly in high-risk individuals. Longitudinal sampling, e.g. 1-2 years apart, to determine the growth rate of clones, would be useful to aid in risk stratification in some individuals, particularly those with rare mutations whose typical growth rate is unknown.

All the risk factors mentioned thus far rely on risk stratification using blood sequencing data. This is not a problem if the individual has already had their blood sequenced, e.g. as part of a cohort study or as part of a solid malignancy work-up, but the majority of individuals in the population, who may be at risk, will not routinely have access to blood sequencing. Until regular whole population DNA sequencing becomes a feasible option, we need to have a way of pre-screening potentially high-risk individuals. Certain red blood cell parameters (red cell distribution width, RDW and mean corpuscle cell volume, MCV) have been found to be associated with high risk[77,259], as is the presence of peripheral blood cytopenias. For individuals with peripheral blood cytopenias without clonal haematopoiesis, termed 'idiopathic cytopenias of uncertain significance' (ICUS), the 5-year probability of progression to myeloid malignancy is ~9%[260,261] . In the setting of clonal haematopoiesis, termed 'clonal cytopenia of unknown significance' (CCUS), this risk increases to 82%[260,261]. Therefore, at present, a simple regular full blood count to identify individuals with an abnormal RDW or MCV seems a reasonable pre-screening test to identify potentially high-risk individuals for sequencing. Given clonal haematopoiesis is not just associated with blood cancer, but also cardiovascular disease, type 2 diabetes, ischemic stroke and overall non-cancer related mortality[3], pursuing the feasibility of screening using population-level DNA sequencing does not seem an unreasonable goal.

Early AML interception

Clonal haematopoiesis clinics already exist in the USA (e.g. MSKCC, Dana-Farber) and are in nascent stages in the UK. With the current lack of a clinically validated tool to predict who is at high-risk of AML, however, current management of these individuals largely consists of counselling and monitoring. The ultimate goal would be to provide high-risk tolerable targeted treatment that could stop pre-AML in its tracks before it progresses. The term 'cancer interception' was first coined in 2011 and refers to the 'detection of pre-cancerous lesions followed by mechanistically based interventions to prevent the formation of cancer'[262].

Whilst historically pharmaceutical companies haven't been interested in pre-malignant conditions, "because people don't die of pre-malignancy"[263,264], they are increasingly recognising their importance. The earlier a cancer is diagnosed and treated, the better the outcome, because the further the cancer (or pre-cancer) has progressed, the greater the genetic complexity and ability to evade therapeutics[263]. A fine balance needs to be struck between treating suspected pre-AML early enough when genetic complexity is low, but late enough to be certain that the individual is at high risk.

Within the past 5 years a number of targeted therapies have emerged which, whilst largely intended for treatment of *de facto* AML, might offer promise in high-risk clonal haematopoiesis. These include IDH1 inhibitors (Ivosidenib), IDH2 inhibitors (Enasidenib)[265], and JAK2 inhibitors (Ruxolitinib, Fedtratinib). Hypomethylating agents have been demonstrated to have particular efficacy against TET2 mutant myeloid malignancies[266,267] and high dose vitamin C supplementation has been shown to limit expansion of TET2-deleted HSPCs[268]. Splicing factor drugs[269,270] and small molecules that bind to TP53 to restore its normal function[271,272] are currently undergoing clinical trials in AML patients. Allele specific inhibitors that trap and inactivate mutant KRAS (G12C) have also been described[273].

Clinical trials are needed to test these therapies in high risk clonal haematopoiesis. The typical long latency before AML develops makes it difficult to determine study end-points, although surrogate end-points such as change in VAF could be used. Great care must be taken to ensure the targeted therapy is not providing a selective pressure that leads to the emergence of more aggressive clones resulting in more rapid progression to malignancy.

6.2 Key unanswered questions

Here, using single-timepoint blood sequencing data amassed from ∼50,000-500,0000 individuals and longitudinal blood samples from 50 pre-AML individuals, we have attempted to map out pre-leukaemic evolution, from acquisition of the first driver mutation through to just before AML diagnosis. We have revealed key evolutionary parameters, inferred which specific variants and mCAs are the fittest and determined the timings of key steps in the evolution to AML. However, several key questions remain unanswered:

What happens in the final stages in the progression to AML?

The closest longitudinal sample we have to AML diagnosis is $\sim$ 3 months which means we are unlikely to have visualised the full evolutionary history preceding AML. We are attempting to obtain surplus bone marrow aspirate slides from the time of AML diagnosis, which will be invaluable for determining the events that trigger the final stage of evolution to AML.

What determines the different patterns of pre-leukaemic evolution?

Why some individuals showed a particular pre-leukaemic evolution pattern and others showed a different one is unknown. Is it due to inter-person variability in overall mutation rates, the specific combination and sequence of mutations acquired or bone marrow microenvironment differences? It is possible that some of the individuals in UKCTOCS developed MDS or MPN before they developed AML and that these are associated with particular evolutionary patterns, although why this would be requires further research. We are attempting to obtain information from NHS Hospital Episode

Statistics (HES) database and the Office of National Statistics (ONS) to determine if MDS or MPN preceding AML in any of the individuals. It is also unclear what triggers 'late' AML to occur. Is a single mutation sufficient to progress to AML and they were just unlucky that this occurred? Are there earlier drivers that we have missed with our targeted approach? Do they have a permissive germline background that results in considerable clonal expansion when a driver mutation occurs? Obtaining the diagnostic bone marrow aspirate slides, performing whole genome sequencing and analysing germline mutations will all shed light on these questions.

What role does the immune system play in clonal dynamics?

The extent to which the immune system shapes pre-leukaemic evolution is also an important area for further research. 'Tumour mass dormancy', where clonal growth and immune-mediated cell death occur at similar rates has been described in other cancers[218,219], and may explain the static evolution observed in samples. Selective pressures exerted by the immune system are also likely to be important. Indeed, recent work in the lung has highlighted the strong selection pressure exerted by the immune microenvironment, producing multiple routes to immune evasion, including loss of heterozygosity in human leucocyte antigens (HLA) and/or depletion of expressed neoantigens[274]. Analysis of 26 tumour types from TCGA data has shown that immune-mediated negative selection acts on MHC-exposed regions of native epitopes[275].

The importance of the 'systemic inflammatory landscape' in clonal haematopoiesis has become increasingly recognised. Individuals with TET2 mutations have been found to have increased circulating levels of IL-8, IL-6 and IL1-β[112,276,277]. Individuals with DNMT3A or ASXL1 mutations also have elevated IL-6 and individuals with SF3B1 have elevated IL-18[277]. Evidence from TET2 knockout mice suggests that dysregulated inflammation (e.g. due to infection or inflammatory disorder) confers a selective advantage to the mutant clone, but that the mutant clones themselves then contribute to the inflammatory milieu, setting in motion a perpetuating cycle of expansion and inflammation[278]. The precise way in which infection and inflammation interact with clonal dynamics is unknown. To investigate this we have set up a small cohort study ('LEGACY': Longitudinal Evaluation of the Growth and Acquisition of Clones over Years) of initially 20 individuals, who will all supply a blood and saliva sample once every 6 weeks, as well as answer a questionnaire related to their health, lifestyle, medications and infections. They will also be provided with a Fitbit to monitor activity levels. By undertaking DNA sequencing, cytokine analysis and immunoprofiling, this study will hopefully shed some light on how health and lifestyle, including infection and inflammation, interact with clonal haematopoiesis.

6.3 Conclusion

Our quantitative analysis of clonal haematopoiesis, combined with an integrated assessment of genetic changes in longitudinal blood samples from individuals who progress to AML reveals important insights into the evolutionary dynamics of mutations in the years preceding AML. As we get closer to understanding which features distinguish pre-malignant from benign clonal evolution it is incumbent that prospective clinical trials of potential tolerable therapeutics continues alongside. With the advent of targeted therapies, the vision of treating high-risk pre-AML may soon become a reality and pre-AML will simply become a chronic disease. In the words of Jan et al, over time, hopefully ‘we will spend less time treating malignant catastrophes and more time preventing them’ [236].

Bibliography

[1] Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197–200 (1975).

[2] Nowell, P.C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).

[3] Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).

[4] Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).

[5] Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).

[6] McKerrell, T., Park, N., Moreno, T., Grove, C.S., Ponstingl, H. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).

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

[8] Acuna-Hidalgo, R., Sengul, H., Steehouwer, M., van de Vorst, M., Vermeulen, S.H. et al. Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *Am. J. Hum. Genet.* **101**, 50–64 (2017).

[9] Zink, F., Stacey, S.N., Norddahl, G.L., Frigge, M.L., Magnusson, O.T. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).

[10] Martincorena, I. & Campbell, P.J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).

[11] Jonason, A.S., Kunala, S., Price, G.J., Restifo, R.J., Spinelli, H.M. et al. Frequent clones of p53-mutated keratinocytes in normal human skin. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14025–14029 (1996).

[12] Ling, G., Persson, A., Berne, B., Uhlén, M., Lundeberg, J. et al. Persistent p53 mutations in single cells from normal human skin. *Am. J. Pathol.* **159**, 1247–1253 (2001).

[13] Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).

[14] Hoang, M.L., Kinde, I., Tomasetti, C., McMahon, K.W., Rosenquist, T.A. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A* **113**, 9846–9851 (2016).

[15] Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719–724 (2009).

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

[17] Jan, M., Snyder, T.M., Corces-Zimmerman, M.R., Vyas, P., Weissman, I.L. et al. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci Transl Med* **4**, 149ra118 (2012).

[18] Welch, J., Ley, T., Link, D., Miller, C., Larson, D. et al. The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**, 264–278 (2012).

[19] UK, C.R.

[20] Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P. et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).

[21] Cancer Genome Atlas Research Network, Ley, T.J., Miller, C., Ding, L., Raphael, B.J. et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).

[22] Dick, J.E. Stem cells: Self-renewal writ in blood. *Nature* **423**, 231–233 (2003).

[23] Weissman, I. Stem cell research: paths to cancer therapies and regenerative medicine. *JAMA* **294**, 1359–1366 (2005).

[24] Corces-Zimmerman, M.R. & Majeti, R. Pre-leukemic evolution of hematopoietic stem cells – the importance of early mutations in leukemogenesis. *Leukemia* **28**, 2276–2282 (2014).

[25] Shouval, R., Shlush, L.I., Yehudai-Resheff, S., Ali, S., Pery, N. et al. Single cell analysis exposes intratumor heterogeneity and suggests that FLT3-ITD is a late event in leukemogenesis. *Exp. Hematol.* **42**, 457–463 (2014).

[26] Shlush, L.I., Zandi, S., Mitchell, A., Chen, W.C., Brandwein, J.M. et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).

[27] Busque, L., Mio, R., Mattioli, J., Brais, E., Blais, N. et al. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* **88**, 59–65 (1996).

[28] Busque, L., Patel, J.P., Figueroa, M., Vasanthakumar, A., Provost, S. et al. Recurrent Somatic TET2 Mutations in Normal Elderly Individuals With Clonal Hematopoiesis. *Nat Genet* **44**, 1179–1181 (2012).

[29] Steensma, D.P., Bejar, R., Jaiswal, S., Lindsley, R.C., Sekeres, M.A. et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9–16 (2015).

[30] Jacobs, K.B., Yeager, M., Zhou, W., Wacholder, S., Wang, Z. et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat Genet* **44**, 651–658 (2012).

[31] Laurie, C.C., Laurie, C.A., Rice, K., Doheny, K.F., Zelnick, L.R. et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat Genet* **44**, 642–650 (2012).

[32] Machiela, M.J., Zhou, W., Sampson, J.N., Dean, M.C., Jacobs, K.B. et al. Characterization of large structural genetic mosaicism in human autosomes. *Am J Hum Genet* **96**, 487–497 (2015).

[33] Terao, C., Suzuki, A., Momozawa, Y., Akiyama, M., Ishigaki, K. et al. Chromosomal alterations among age-related hematopoietic clones in Japan. *Nature* **584**, 130–135 (2020).

[34] Loh, P.R., Genovese, G., Handsaker, R.E., Finucane, H.K., Reshef, Y.A. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).

[35] Loh, P.R., Genovese, G. & McCarroll, S.A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).

[36] Niroula, A., Sekar, A., Murakami, M.A., Trinder, M., Agrawal, M. et al. Distinction of lymphoid and myeloid clonal hematopoiesis. *Nat Med* **27**, 1921–1927 (2021).

[37] Abelson, S. & Wang, J.C.Y. Age-related clonal hematopoiesis: implications for hematopoietic stem cell transplantation. *Current Opinion in Hematology* **25**, 441 (2018).

[38] Desai, P., Mencia-Trinchant, N., Savenkov, O., Simon, M.S., Cheang, G. et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nature Medicine* **24**, 1015–1023 (2018).

[39] Young, A.L., Tong, R.S., Birmann, B.M. & Druley, T.E. Clonal haematopoiesis and risk of acute myeloid leukemia. *Haematologica* [Epub ahead of print] (2019).

[40] Dittmer, D., Pati, S., Zambetti, G., Chu, S., Teresky, A.K. et al. Gain of function mutations in p53. *Nat Genet* **4**, 42–46 (1993).

[41] Olive, K.P., Tuveson, D.A., Ruhe, Z.C., Yin, B., Willis, N.A. et al. Mutant p53 gain of function in two mouse models of Li-Fraumeni syndrome. *Cell* **119**, 847–860 (2004).

[42] Boettcher, S., Miller, P.G., Sharma, R., McConkey, M., Leventhal, M. et al. A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies. *Science* **365**, 599–604 (2019).

[43] Clayton, E., Doupé, D.P., Klein, A.M., Winton, D.J., Simons, B.D. et al. A single type of progenitor cell maintains normal epidermis. *Nature* **446**, 185–189 (2007).

[44] Klein, A.M., Doupé, D.P., Jones, P.H. & Simons, B.D. Kinetics of cell division in epidermal maintenance. *Phys. Rev. E* **76**, 021910 (2007). Publisher: American Physical Society.

[45] Desai, M.M. & Fisher, D.S. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759–1798 (2007).

[46] Simons, B.D. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. *Proc. Natl. Acad. Sci. USA* **113**, 128–133 (2016).

[47] Holstege, H., Pfeiffer, W., Sie, D., Hulsman, M., Nicholas, T.J. 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).

[48] Lee-Six, H., Øbro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).

[49] Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

[50] Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).

[51] Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).

[52] Hwang, D.G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13994–14001 (2004).

[53] Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).

[54] Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).

[55] Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756–766 (2011).

[56] Park, C., Qian, W. & Zhang, J. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep.* **13**, 1123–1129 (2012).

[57] Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

[58] Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A. et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).

[59] Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J. et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).

[60] Buescher, E.S., Alling, D.W. & Gallin, J.I. Use of an X-linked human neutrophil marker to estimate timing of lyonization and size of the dividing stem cell pool. *J. Clin. Invest.* **76**, 1581–1584 (1985).

[61] 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**, 2665–2667 (2002).

[62] Wang, J.C., Doedens, M. & Dick, J.E. Primitive human hematopoietic cells are enriched in cord blood compared with adult bone marrow or mobilized peripheral blood as measured by the quantitative in vivo SCID-repopulating cell assay. *Blood* **89**, 3919–3924 (1997).

[63] Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genet.* **48**, 1193–1203 (2016).

[64] 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. USA* **96**, 3120–3125 (1999).

[65] Dingli, D. & Pacheco, J.M. Allometric scaling of the active hematopoietic stem cell pool across mammals. *PLoS ONE* **1**, e2 (2006).

[66] 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).

[67] Rufer, N., Brümmendorf, T.H., Kolvraa, S., Bischoff, C., Christensen, K. 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. *The J. Exp. Med.* **190**, 157–168 (1999).

[68] Grove, C.S. & Vassiliou, G.S. Acute myeloid leukaemia: a paradigm for the clonal evolution of cancer? *Dis Model Mech* **7**, 941–951 (2014).

[69] Kinzler, K.W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).

[70] Heyer, J., Kwong, L.N., Lowe, S.W. & Chin, L. Non-germline genetically engineered mouse models for translational cancer research. *Nat. Rev. Cancer* **10**, 470–480 (2010).

[71] Vassiliou, G.S., Cooper, J.L., Rad, R., Li, J., Rice, S. et al. Mutant nucleophosmin and cooperating pathways drive leukemia initiation and progression in mice. *Nat. Genet.* **43**, 470–475 (2011).

[72] Ortmann, C.A., Kent, D.G., Nangalia, J., Silber, Y., Wedge, D.C. et al. Effect of mutation order on myeloproliferative neoplasms. *N. Engl. J. Med.* **372**, 601–612 (2015).

[73] Wiemels, J.L., Xiao, Z., Buffler, P.A., Maia, A.T., Ma, X. et al. In utero origin of t(8;21) AML1-ETO translocations in childhood acute myeloid leukemia. *Blood* **99**, 3801–3805 (2002).

[74] Afshar-Sterle, S., Zotos, D., Bernard, N.J., Scherger, A.K., Rödling, L. et al. Fas ligand-mediated immune surveillance by T cells is essential for the control of spontaneous B cell lymphomas. *Nat. Med.* **20**, 283–290 (2014).

[75] Coombs, C.C., Zehir, A., Devlin, S.M., Kishtagari, A., Syed, A. et al. Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. *Cell Stem Cell* **21**, 374–382.e4 (2017).

[76] Desai, P., Mencia-Trinchant, N., Savenkov, O., Simon, M.S., Cheang, G. et al. Somatic Mutations Predict Acute Myeloid Leukemia Years Before Diagnosis. *bioRxiv* 237941 (2017).

[77] Abelson, S., Collord, G., Ng, S.W.K., Weissbrod, O., Cohen, N.M. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* 1 (2018).

[78] Zon, L.I. Intrinsic and extrinsic control of haematopoietic stem-cell self-renewal. *Nature* **453**, 306–313 (2008).

[79] Bolton, K.L., Ptashkin, R.N., Gao, T., Braunstein, L., Devlin, S.M. et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nat Genet* **52**, 1219–1226 (2020).

[80] Ho, A.D. & Wagner, W. The beauty of asymmetry: asymmetric divisions and self-renewal in the haematopoietic system. *Curr. Opin. Hematol.* **14**, 330–336 (2007).

[81] Till, J.E., McCulloch, E.A. & Siminovitch, L. A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proc Natl Acad Sci U S A* **51**, 29–36 (1964).

[82] Zhang, Y.V., Cheong, J., Ciapurin, N., McDermitt, D.J. & Tumbar, T. Distinct self-renewal and differentiation phases in the niche of infrequently dividing hair follicle stem cells. *Cell Stem Cell* **5**, 267–278 (2009).

[83] Loeffler, M. & Roeder, I. Tissue stem cells: definition, plasticity, heterogeneity, self-organization and models–a conceptual approach. *Cells Tissues Organs (Print)* **171**, 8–26 (2002).

[84] Ogawa, M. Differentiation and proliferation of hematopoietic stem cells. *Blood* **81**, 2844–2853 (1993).

[85] Lopez-Garcia, C., Klein, A.M., Simons, B.D. & Winton, D.J. Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330**, 822–825 (2010).

[86] Boulais, P.E. & Frenette, P.S. Making sense of hematopoietic stem cell niches. *Blood* **125**, 2621–2629 (2015).

[87] Mesa, K.R., Rompolas, P. & Greco, V. The Dynamic Duo: Niche/Stem Cell Interdependency. *Stem Cell Reports* **4**, 961–966 (2015).

[88] Guilak, F., Cohen, D.M., Estes, B.T., Gimble, J.M., Liedtke, W. et al. Control of stem cell fate by physical interactions with the extracellular matrix. *Cell Stem Cell* **5**, 17–26 (2009).

[89] Metcalf, D. Lineage commitment in the progeny of murine hematopoietic preprogenitor cells: influence of thrombopoietin and interleukin 5. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6408–6412 (1998).

[90] Metcalf, D. Lineage commitment of hemopoietic progenitor cells in developing blast cell colonies: influence of colony-stimulating factors. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 11310–11314 (1991).

[91] Hwang, S.M., Im, K., Chang, Y.H., Park, H.S., Kim, J.A. 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. *Leuk. Res.* **71**, 92–94 (2018).

[92] Mitchell, E., Chapman, M.S., Williams, N., Dawson, K., Mende, N. et al. Clonal dynamics of haematopoiesis across the human lifespan. Tech. rep. (2021). Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results.

[93] Watson, C.J., Papula, A.L., Poon, G.Y.P., Wong, W.H., Young, A.L. et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).

[94] 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).

[95] Yoda, A., Adelmant, G., Tamburini, J., Chapuy, B., Shindoh, N. et al. GNB1 activating mutations promote myeloid and lymphoid neoplasms targetable by combined PI3K/mTOR inhibition. *Blood* **124**, 3567–3567 (2014).

[96] Wong, T.N., Ramsingh, G., Young, A.L., Miller, C.A., Touma, W. et al. The role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature* **518**, 552–555 (2015).

[97] Hsu, J.I., Dayaram, T., Tovy, A., De Braekeleer, E., Jeong, M. et al. PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. *Cell Stem Cell* **23**, 700–713.e6 (2018).

[98] Bolton, K. Oncologic therapy for solid tumors alters the risk of clonal hematopoiesis. ASH (2018).

[99] Takizawa, H., Boettcher, S. & Manz, M.G. Demand-adapted regulation of early hematopoiesis in infection and inflammation. *Blood* **119**, 2991–3002 (2012).

[100] Meisel, M., Hinterleitner, R., Pacis, A., Chen, L., Earley, Z.M. et al. Microbial signals drive pre-leukaemic myeloproliferation in a Tet2 -deficient host. *Nature* **557**, 580 (2018).

[101] King, K.Y. & Goodell, M.A. Inflammatory modulation of HSCs: viewing the HSC as a foundation for the immune response. *Nat. Rev. Immunol.* **11**, 685–692 (2011).

[102] Murai, K., Skrupskelyte, G., Piedrafita, G., Hall, M., Kostiou, V. et al. Epidermal Tissue Adapts to Restrain Progenitors Carrying Clonal p53 Mutations. *Cell Stem Cell* **23**, 687–699.e8 (2018).

[103] Fabre, M.A., Almeida, J.G.d., Fiorillo, E., Mitchell, E., Damaskou, A. et al. The longitudinal dynamics and natural history of clonal haematopoiesis. Tech. rep. (2021).

[104] Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).

[105] Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).

[106] Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).

[107] Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature Genet.* **48**, 238–244 (2016).

[108] Williams, M.J., Werner, B., Heide, T., Curtis, C., Barnes, C.P. et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genet.* **50**, 895 (2018).

[109] Nielsen, C., Bojesen, S.E., Nordestgaard, B.G., Kofoed, K.F. & Birgens, H.S. JAK2v617f somatic mutation in the general population: myeloproliferative neoplasm development and progression rate. *Haematologica* **99**, 1448–1455 (2014).

[110] Jeong, M., Park, H.J., Celik, H., Ostrander, E.L., Reyes, J.M. et al. Loss of Dnmt3a Immortalizes Hematopoietic Stem Cells In Vivo. *Cell Rep.* **23**, 1–10 (2018).

[111] Chan, S.M. & Majeti, R. Role of DNMT3a, TET2, and IDH1/2 mutations in pre-leukemic stem cells in acute myeloid leukemia. *Int J Hematol* **98**, 648–657 (2013).

[112] Jaiswal, S., Natarajan, P., Silver, A.J., Gibson, C.J., Bick, A.G. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med.* **377**, 111–121 (2017).

[113] Liu, J., Chen, X., Wang, J., Zhou, S., Wang, C.L. et al. Biological background of the genomic variations of cf-DNA in healthy individuals. *Ann. Oncol.* [Epub ahead of print] (2019).

[114] Swanton, C., Venn, O., Aravanis, A., Hubbell, E., Maddala, T. et al. Prevalence of clonal hematopoiesis of indeterminate potential (CHIP) measured by an ultra-sensitive sequencing assay: Exploratory analysis of the Circulating Cancer Genome Atlas (CCGA) study. *J. Clin. Oncol.* **36**, suppl; abstr 12003 (2018).

[115] Yoshizato, T., Dumitriu, B., Hosokawa, K., Makishima, H., Yoshida, K. et al. Somatic mutations and clonal hematopoiesis in aplastic anemia. *N. Engl. J. Med.* **373**, 35–47 (2015).

[116] Takahashi, K., Wang, F., Kantarjian, H., Doss, D., Khanna, K. et al. Preleukaemic clonal haemopoiesis and risk of therapy-related myeloid neoplasms: a case-control study. *Lancet Oncol.* **18**, 100–111 (2017).

[117] Bolton, K.L., Gillis, N.K., Coombs, C.C., Takahashi, K., Zehir, A. et al. Managing clonal hematopoiesis in patients with solid tumors. *J. Clin. Oncol.* JCO.18.00331 (2018).

[118] Wong, T.N., Miller, C.A., Jotte, M.R.M., Bagegni, N., Baty, J.D. et al. Cellular stressors contribute to the expansion of hematopoietic clones of varying leukemic potential. *Nat. Commun.* **9**, 455 (2018).

[119] Lindsley, R.C., Saber, W., Mar, B.G., Redd, R., Wang, T. et al. Prognostic mutations in myelodysplastic syndrome after stem-cell transplantation. *N. Engl. J. Med.* **376**, 536–547 (2017).

[120] Bullinger, L., Krönke, J., Schön, C., Radtke, I., Urlbauer, K. et al. Identification of acquired copy number alterations and uniparental disomies in cytogenetically normal acute myeloid leukemia using high-resolution single-nucleotide polymorphism analysis. *Leukemia* **24**, 438–449 (2010).

[121] Forsberg, L.A., Rasi, C., Razzaghian, H.R., Pakalapati, G., Waite, L. et al. Age-related somatic structural changes in the nuclear genome of human blood cells. *Am J Hum Genet* **90**, 217–228 (2012).

[122] Gao, T., Ptashkin, R., Bolton, K.L., Sirenko, M., Fong, C. et al. Interplay between chromosomal alterations and gene mutations shapes the evolutionary trajectory of clonal hematopoiesis. *Nat Commun* **12**, 338 (2021).

[123] Germing, U., Lauseker, M., Hildebrandt, B., Symeonidis, A., Cermak, J. et al. Survival, prognostic factors and rates of leukemic transformation in 381 untreated patients with MDS and del(5q): a multicenter study. *Leukemia* **26**, 1286–1292 (2012).

[124] Lee-Six, H., Øbro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).

[125] Osorio, F.G., Rosendahl Huber, A., Oka, R., Verheul, M., Patel, S.H. et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Rep.* **25**, 2308–2316.e4 (2018).

[126] Grimwade, D., Walker, H., Oliver, F., Wheatley, K., Harrison, C. et al. The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children's Leukaemia Working Parties. *Blood* **92**, 2322–2333 (1998).

[127] Tasaka, T., Tohyama, K., Kishimoto, M., Ohyashiki, K., Mitani, K. et al. Myelodysplastic syndrome with chromosome 5 abnormalities: a nationwide survey in Japan. *Leukemia* **22**, 1874–1881 (2008).

[128] McGhee, E.M., Cohen, N.R., Wolf, J.L., Ledesma, C.T. & Cotter, P.D. Monosomy 16 as the sole abnormality in myeloid malignancies. *Cancer Genet Cytogenet* **118**, 163–166 (2000).

[129] Lin, S.H., Brown, D.W., Rose, B., Day, F., Lee, O.W. et al. Incident disease associations with mosaic chromosomal alterations on autosomes, X and Y chromosomes: insights from a phenome-wide association study in the UK Biobank. *Cell Biosci* **11**, 143 (2021).

[130] Shahrabi, S., Khodadi, E., Saba, F., Shahjahani, M. & Saki, N. Sex chromosome changes in leukemia: cytogenetics and molecular aspects. *Hematology* **23**, 139–147 (2018).

[131] Ouseph, M.M., Hasserjian, R.P., Dal Cin, P., Lovitch, S.B., Steensma, D.P. et al. Genomic alterations in patients with somatic loss of the Y chromosome as the sole cytogenetic finding in bone marrow cells. *Haematologica* **106**, 555–564 (2021).

[132] Jacobs, I.J., Menon, U., Ryan, A., Gentry-Maharaj, A., Burnell, M. et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *The Lancet* **387**, 945–956 (2016).

[133] Widschwendter, M., Zikan, M., Wahl, B., Lempiäinen, H., Paprotka, T. et al. The potential of circulating tumor DNA methylation analysis for the early detection and management of ovarian cancer. *Genome Medicine* **9**, 116 (2017).

[134] Fox, E.J., Reid-Bayliss, K.S., Emond, M.J. & Loeb, L.A. Accuracy of Next Generation Sequencing Platforms. *Next Gener Seq Appl* **1** (2014).

[135] Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics* **3** (2021).

[136] Hiatt, J.B., Pritchard, C.C., Salipante, S.J., O'Roak, B.J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23**, 843–854 (2013).

[137] Kennedy, S.R., Schmitt, M.W., Fox, E.J., Kohrn, B.F., Salk, J.J. et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature Protocols* **9**, 2586–2606 (2014).

[138] Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).

[139] Wong, W.H., Tong, R.S., Young, A.L. & Druley, T.E. Rare Event Detection Using Error-corrected DNA and RNA Sequencing. *J Vis Exp* (2018).

[140] Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B. et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14508–14513 (2012).

[141] Abascal, F., Harvey, L.M.R., Mitchell, E., Lawson, A.R.J., Lensing, S.V. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).

[142] The Cancer Genome Atlas Research Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N Engl J Med* **368**, 2059–2074 (2013).

[143] Kihara, R., Nagata, Y., Kiyoi, H., Kato, T., Yamamoto, E. et al. Comprehensive analysis of genetic alterations and their prognostic impacts in adult acute myeloid leukemia patients. *Leukemia* **28**, 1586–1595 (2014).

[144] Arber, D.A., Bruning, R., Le Beau, M.M., Falini, B., Vardiman, J. et al. Acute myeloid leukaemia and related precursor neoplasms. In *WHO classification of tumours of haematopoietic and lymphoid tissues*, 130–171. IARC Press, Lyon, France, revised 4th ed. ed. (2017).

[145] Gondek, L.P., Tiu, R., O'Keefe, C.L., Sekeres, M.A., Theil, K.S. et al. Chromosomal lesions and uniparental disomy detected by SNP arrays in MDS, MDS/MPD, and MDS-derived AML. *Blood* **111**, 1534–1542 (2008).

[146] Sebaa, A., Ades, L., Baran-Marzack, F., Mozziconacci, M.J., Penther, D. et al. Incidence of 17p deletions and TP53 mutation in myelodysplastic syndrome and acute myeloid leukemia with 5q deletion. *Genes, Chromosomes and Cancer* **51**, 1086–1092 (2012).

[147] Gronseth, C.M., McElhone, S.E., Storer, B.E., Kroeger, K.A., Sandhu, V. et al. Prognostic significance of acquired copy-neutral loss of heterozygosity in acute myeloid leukemia. *Cancer* **121**, 2900–2908 (2015).

[148] Xu, X., Johnson, E.B., Leverton, L., Arthur, A., Watson, Q. et al. The advantage of using SNP array in clinical testing for hematological malignancies—a comparative study of three genetic testing methods. *Cancer Genetics* **206**, 317–326 (2013).

[149] Shen, W., Szankasi, P., Sederberg, M., Schumacher, J., Frizzell, K.A. et al. Concurrent detection of targeted copy number variants and mutations using a myeloid malignancy next generation sequencing panel allows comprehensive genetic analysis using a single testing strategy. *Br J Haematol* **173**, 49–58 (2016).

[150] Shen, W., Szankasi, P., Durtschi, J., Kelley, T.W. & Xu, X. Genome-Wide Copy Number Variation Detection Using NGS: Data Analysis and Interpretation. In S.S. Murray (editor), *Tumor Profiling: Methods and Protocols*, Methods in Molecular Biology, 113–124. Springer, New York, NY (2019).

[151] McKerrell, T., Moreno, T., Ponstingl, H., Bolli, N., Dias, J.M.L. et al. Development and validation of a comprehensive genomic diagnostic tool for myeloid malignancies. *Blood* **128**, e1–e9 (2016).

[152] Rowley, J.D. Identificaton of a translocation with quinacrine fluorescence in a patient with acute leukemia. *Ann Genet* **16**, 109–112 (1973).

[153] Rowley, J.D., Golomb, H.M. & Dougherty, C. 15/17 translocation, a consistent chromosomal change in acute promyelocytic leukaemia. *Lancet* **1**, 549–550 (1977).

[154] Mitelman, F., Johansson, B. & Mertens, F. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2021).

[155] Speck, N.A. & Gilliland, D.G. Core-binding factors in haematopoiesis and leukaemia. *Nat Rev Cancer* **2**, 502–513 (2002).

[156] Gale, K.B., Ford, A.M., Repp, R., Borkhardt, A., Keller, C. et al. Backtracking leukemia to birth: Identification of clonotypic gene fusion sequences in neonatal bloodspots. *Proc Natl Acad Sci U S A* **94**, 13950–13954 (1997).

[157] Wiemels, J.L., Alexander, F.E., Cazzaniga, G., Biondi, A., Mayer, S.P. et al. Microclustering of TEL-AML1 translocation breakpoints in childhood acute lymphoblastic leukemia. *Genes, Chromosomes and Cancer* **29**, 219–228 (2000).

[158] Basecke, J., Cepek, L., Mannhalter, C., Krauter, J., Hildenhagen, S. et al. Transcription of AML1/ETO in bone marrow and cord blood of individuals without acute myelogenous leukemia. *Blood* **100**, 2267 (2002).

[159] Bose, S., Deininger, M., Gora-Tybor, J., Goldman, J.M. & Melo, J.V. The Presence of Typical and Atypical BCR-ABL Fusion Genes in Leukocytes of Normal Individuals: Biologic Significance and Implications for the Assessment of Minimal Residual Disease. *Blood* **92**, 3362–3367 (1998).

[160] Janz, S., Potter, M. & Rabkin, C.S. Lymphoma- and leukemia-associated chromosomal translocations in healthy individuals. *Genes, Chromosomes and Cancer* **36**, 211–223 (2003).

[161] Welch, J.S., Westervelt, P., Ding, L., Larson, D.E., Klco, J.M. et al. Use of whole genome sequencing to diagnose a cryptic fusion oncogene. *JAMA* **305**, 1577–1584 (2011).

[162] Duncavage, E.J., Abel, H.J., Szankasi, P., Kelley, T.W. & Pfeifer, J.D. Targeted next generation sequencing of clinically significant gene mutations and translocations in leukemia. *Mod Pathol* **25**, 795–804 (2012).

[163] Abel, H.J., Al-Kateb, H., Cottrell, C.E., Bredemeyer, A.J., Pritchard, C.C. et al. Detection of Gene Rearrangements in Targeted Clinical Next-Generation Sequencing. *J Mol Diagn* **16**, 405–417 (2014).

[164] Prieto-Conde, M.I., Corchete, L.A., García-Álvarez, M., Jiménez, C., Medina, A. et al. A New Next-Generation Sequencing Strategy for the Simultaneous Analysis of Mutations and Chromosomal Rearrangements at DNA Level in Acute Myeloid Leukemia Patients. *The Journal of Molecular Diagnostics* **22**, 60–71 (2020).

[165] Meyer, C., Burmeister, T., Gröger, D., Tsaur, G., Fechina, L. et al. The MLL recombinome of acute leukemias in 2017. *Leukemia* **32**, 273–284 (2018).

[166] Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941–D947 (2019).

[167] Polprasert, C., Schulze, I., Sekeres, M.A., Makishima, H., Przychodzen, B. et al. Inherited and Somatic Defects in DDX41 in Myeloid Neoplasms. *Cancer Cell* **27**, 658–670 (2015).

[168] Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[169] Dale, R.K., Pedersen, B.S. & Quinlan, A.R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).

[170] Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

[171] Harper, D.P. & Aplan, P.D. Chromosomal Rearrangements Leading to MLL Gene Fusions: Clinical and Biological Aspects. *Cancer Res* **68**, 10024–10027 (2008).

[172] Basecke, J., Whelan, J.T., Griesinger, F. & Bertrand, F.E. The MLL partial tandem duplication in acute myeloid leukaemia. *Br J Haematol* **135**, 438–449 (2006).

[173] Döhner, K., Tobis, K., Ulrich, R., Fröhling, S., Benner, A. et al. Prognostic significance of partial tandem duplications of the MLL gene in adult patients 16 to 60 years old with acute myeloid leukemia and normal cytogenetics: a study of the Acute Myeloid Leukemia Study Group Ulm. *J Clin Oncol* **20**, 3254–3261 (2002).

[174] Garçon, L., Libura, M., Delabesse, E., Valensi, F., Asnafi, V. et al. DEK-CAN molecular monitoring of myeloid malignancies could aid therapeutic stratification. *Leukemia* **19**, 1338–1344 (2005).

[175] Monma, F., Nishii, K., Shiga, J., Sugahara, H., Lorenzo, F. et al. Detection of the CBFB/MYH11 fusion gene in de novo acute myeloid leukemia (AML): A single-institution study of 224 Japanese AML patients. *Leukemia Research* **31**, 471–476 (2007).

[176] Konoplev, S. & Bueso-Ramos, C. Acute Myeloid Leukaemias with Recurrent Cytogenetic Abnormalities. In *Molecular Pathology of Hematolymphoid Disease*, 428–448. Springer (2010).

[177] Wang, H.Y. & Rashidi, H.H. The New Clinicopathologic and Molecular Findings in Myeloid Neoplasms With inv(3)(q21q26)/t(3;3)(q21;q26.2). *Archives of Pathology & Laboratory Medicine* **140**, 1404–1410 (2016).

[178] Broad Institute. Picard tools. http://broadinstitute.github.io/picard/ (version 2.18.15).

[179] Fulcrum Genomics. fgbio: https://github.com/fulcrumgenomics/fgbio.

[180] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* (2013). ArXiv: 1303.3997.

[181] Van der Auwera, G. & O'Connor, B. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media, 1st edition ed. (2020).

[182] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

[183] Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O. et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* **44**, e108 (2016).

[184] Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).

[185] Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).

[186] Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M. et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* **45**, D840–D845 (2017).

[187] Salk, J.J., Schmitt, M.W. & Loeb, L.A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat Rev Genet* **19**, 269–285 (2018).

[188] You, X., Thiruppathi, S., Liu, W., Cao, Y., Naito, M. et al. Detection of genome-wide low-frequency mutations with Paired-End and Complementary Consensus Sequencing (PECC-Seq) revealed end-repair-derived artifacts as residual errors. *Arch Toxicol* **94**, 3475–3485 (2020).

[189] Loh, P.R., Palamara, P.F. & Price, A.L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* **48**, 811–816 (2016).

[190] Heger, A. & Jacobs, K. pysam v0.16.0.1: https://github.com/pysam-developers/pysam.

[191] Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M. et al. BreakDancer: An algorithm for high resolution mapping of genomic structural variation. *Nat Methods* **6**, 677–681 (2009).

[192] Wang, J., Mullighan, C.G., Easton, J., Roberts, S., Ma, J. et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* **8**, 652–654 (2011).

[193] Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

[194] Layer, R.M., Chiang, C., Quinlan, A.R. & Hall, I.M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology* **15**, R84 (2014).

[195] Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

[196] Cameron, D.L., Schröder, J., Penington, J.S., Do, H., Molania, R. et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res* **27**, 2050–2060 (2017).

[197] Gong, T., Hayes, V.M. & Chan, E.K.F. Detection of somatic structural variants from short-read next-generation sequencing data. *Brief Bioinform* **22**, bbaa056 (2020).

[198] Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192 (2013).

[199] Broad Institute. Decoding SAM flags: https://broadinstitute.github.io/picard/explain-flags.html.

[200] Gerstung, M., Papaemmanuil, E. & Campbell, P.J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198–1204 (2014).

[201] Society, A.C. Cancer Facts & Figures 2021. *Atlanta: American Cancer Society* (2021).

[202] McKerrell, T. & Vassiliou, G.S. Aging as a driver of leukemogenesis. *Sci. Transl. Med.* **7**, 306fs38 (2015).

[203] Awada, H., Nagata, Y., Goyal, A., Asad, M.F., Patel, B. et al. Invariant phenotype and molecular association of biallelic TET2 mutant myeloid neoplasia. *Blood Adv* **3**, 339–349 (2019).

[204] Ho, P.A., Zeng, R., Alonzo, T.A., Gerbing, R.B., Miller, K.L. et al. Prevalence and prognostic implications of WT1 mutations in pediatric acute myeloid leukemia (AML): a report from the Children's Oncology Group. *Blood* **116**, 702–710 (2010).

[205] Yang, L., Rau, R. & Goodell, M.A. DNMT3A in haematological malignancies. *Nat Rev Cancer* **15**, 152–165 (2015).

[206] Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T. et al. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat Commun* **11**, 5327 (2020).

[207] Miles, L.A., Bowman, R.L., Merlinsky, T.R., Csete, I.S., Ooi, A.T. et al. Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* **587**, 477–482 (2020).

[208] Fearon, E.R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).

[209] Mroz, E.A., Tward, A.D., Pickering, C.R., Myers, J.N., Ferris, R.L. et al. High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer* **119**, 3034–3042 (2013).

[210] Cooke, S.L., Temple, J., Macarthur, S., Zahra, M.A., Tan, L.T. et al. Intra-tumour genetic heterogeneity and poor chemoradiotherapy response in cervical cancer. *Br J Cancer* **104**, 361–368 (2011).

[211] Itzykson, R., Duployez, N., Fasan, A., Decool, G., Marceau-Renaut, A. et al. Clonal interference of signaling mutations worsens prognosis in core-binding factor acute myeloid leukemia. *Blood* **132**, 187–196 (2018).

[212] Gerrish, P.J. & Lenski, R.E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102**, 127 (1998).

[213] Gao, R., Davis, A., McDonald, T.O., Sei, E., Shi, X. et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet* **48**, 1119–1130 (2016).

[214] Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A. et al. Punctuated Evolution of Prostate Cancer Genomes. *Cell* **153**, 666–677 (2013).

[215] Maher, C.A. & Wilson, R.K. Chromothripsis and Human Disease: Piecing Together the Shattering Process. *Cell* **148**, 29–32 (2012).

[216] Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R. et al. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144**, 27–40 (2011). Publisher: Elsevier.

[217] Li, X., Liu, D., Zhang, L., Wang, H., Li, Y. et al. The comprehensive DNA methylation landscape of hematopoietic stem cell development. *Cell Discov* **7**, 86 (2021).

[218] Endo, H. & Inoue, M. Dormancy in cancer. *Cancer Sci* **110**, 474–480 (2019).

[219] Aguirre-Ghiso, J.A. Models, mechanisms and clinical evidence for cancer dormancy. *Nat Rev Cancer* **7**, 834–846 (2007).

[220] Zhang, B., Nguyen, L.X.T., Li, L., Zhao, D., Kumar, B. et al. Bone marrow niche trafficking of miR-126 controls the self-renewal of leukemia stem cells in chronic myelogenous leukemia. *Nat Med* **24**, 450–462 (2018).

[221] Agarwal, P., Isringhausen, S., Li, H., Paterson, A.J., He, J. et al. Mesenchymal niche-specific expression of Cxcl12 controls quiescence of treatment-resistant leukemia stem cells. *Cell Stem Cell* **24**, 769–784.e6 (2019).

[222] Senft, D. & Jeremias, I. Tumor Cell Dormancy—Triggered by the Niche. *Developmental Cell* **49**, 311–312 (2019).

[223] Medyouf, H. The microenvironment in human myeloid malignancies: emerging concepts and therapeutic implications. *Blood* **129**, 1617–1626 (2017).

[224] Büchner, T., Berdel, W.E., Haferlach, C., Haferlach, T., Schnittger, S. et al. Age-related risk profile and chemotherapy dose response in acute myeloid leukemia: a study by the German Acute Myeloid Leukemia Cooperative Group. *J Clin Oncol* **27**, 61–69 (2009).

[225] Rao, A.V., Valk, P.J.M., Metzeler, K.H., Acharya, C.R., Tuchman, S.A. et al. Age-specific differences in oncogenic pathway dysregulation and anthracycline sensitivity in patients with acute myeloid leukemia. *J Clin Oncol* **27**, 5580–5586 (2009).

[226] Silva, P., Neumann, M., Schroeder, M.P., Vosberg, S., Schlee, C. et al. Acute myeloid leukemia in the elderly is characterized by a distinct genetic and epigenetic landscape. *Leukemia* **31**, 1640–1644 (2017).

[227] De-Morgan, A., Meggendorfer, M., Haferlach, C. & Shlush, L. Male predominance in AML is associated with specific preleukemic mutations. *Leukemia* **35**, 867–870 (2021).

[228] Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D. et al. The Life History of 21 Breast Cancers. *Cell* **149**, 994–1007 (2012).

[229] Miller, C.A., White, B.S., Dees, N.D., Griffith, M., Welch, J.S. et al. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLOS Computational Biology* **10**, e1003665 (2014).

[230] Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11**, 396–398 (2014).

[231] Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R.B. et al. Fast and scalable inference of multi-sample cancer lineages. *Genome Biology* **16**, 91 (2015).

[232] Niknafs, N., Beleva-Guthrie, V., Naiman, D.Q. & Karchin, R. SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing. *PLOS Computational Biology* **11**, e1004416 (2015).

[233] Zare, H., Wang, J., Hu, A., Weber, K., Smith, J. et al. Inferring Clonal Composition from Multiple Sections of a Breast Cancer. *PLOS Computational Biology* **10**, e1003703 (2014).

[234] Lindsley, R.C., Mar, B.G., Mazzola, E., Grauman, P.V., Shareef, S. et al. Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. *Blood* **125**, 1367–1376 (2015).

[235] Fairbanks, D.J. Mendel and Darwin: untangling a persistent enigma. *Heredity* **124**, 263–273 (2020).

[236] Jan, M., Ebert, B.L. & Jaiswal, S. Clonal hematopoiesis. *Seminars in Hematology* (2016).

[237] Griffith, J.F., Yeung, D.K.W., Ma, H.T., Leung, J.C.S., Kwok, T.C.Y. et al. Bone marrow fat content in the elderly: a reversal of sex difference seen in younger subjects. *J Magn Reson Imaging* **36**, 225–230 (2012).

[238] Roerink, S.F., Sasaki, N., Lee-Six, H., Young, M.D., Alexandrov, L.B. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457–462 (2018).

[239] Petljak, M., Alexandrov, L.B., Brammeld, J.S., Price, S., Wedge, D.C. et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).

[240] Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

[241] Beerman, I., Bhattacharya, D., Zandi, S., Sigvardsson, M., Weissman, I.L. et al. Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. *PNAS* **107**, 5465–5470 (2010). ISBN: 9781000834109 Publisher: National Academy of Sciences Section: Biological Sciences.

[242] Ogawa, T., Kitagawa, M. & Hirokawa, K. Age-related changes of human bone marrow: a histometric estimation of proliferative cells, apoptotic cells, T cells, B cells and macrophages. *Mech Ageing Dev* **117**, 57–68 (2000).

[243] Linton, P.J. & Dorshkind, K. Age-related changes in lymphocyte development and function. *Nat Immunol* **5**, 133–139 (2004).

[244] Guralnik, J.M., Eisenstaedt, R.S., Ferrucci, L., Klein, H.G. & Woodman, R.C. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. *Blood* **104**, 2263–2268 (2004).

[245] Kuranda, K., Vargaftig, J., de la Rochere, P., Dosquet, C., Charron, D. et al. Age-related changes in human hematopoietic stem/progenitor cells. *Aging Cell* **10**, 542–546 (2011).

[246] Pang, W.W., Price, E.A., Sahoo, D., Beerman, I., Maloney, W.J. 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).

[247] Henry, C.J., Marusyk, A., Zaberezhnyy, V., Adane, B. & DeGregori, J. Declining lymphoid progenitor fitness promotes aging-associated leukemogenesis. *Proc Natl Acad Sci U S A* **107**, 21713–21718 (2010).

[248] Marusyk, A. & DeGregori, J. Declining cellular fitness with age promotes cancer initiation by selecting for adaptive oncogenic mutations. *Biochim Biophys Acta* **1785**, 1–11 (2008).

[249] Lawson, A.R.J., Abascal, F., Coorens, T.H.H., Hooks, Y., O'Neill, L. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).

[250] Sanders, M.A., Chew, E., Flensburg, C., Zeilemaker, A., Miller, S.E. et al. Germline loss of MBD4 predisposes to leukaemia due to a mutagenic cascade driven by 5mc. *bioRxiv* 180588 (2017).

[251] Kusne, Y., Lasho, T., Mangaonkar, A., Tefferi, A., Gangat, N. et al. Remarkable stability in clonal hematopoiesis involving leukemia-driver genes in patients without underlying myeloid neoplasms. *Am J Hematol* **96**, E392–E396 (2021).

[252] Bissell, M.J. & Hines, W.C. Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression. *Nat Med* **17**, 320–329 (2011).

[253] Arends, C.M., Galan-Sousa, J., Hoyer, K., Chan, W., Jäger, M. et al. Hematopoietic lineage distribution and evolutionary dynamics of clonal hematopoiesis. *Leukemia* **32**, 1908–1919 (2018).

[254] Nishanth, G., Wolleschak, D., Fahldieck, C., Fischer, T., Mullally, A. et al. Gain of function in Jak2V617F-positive T-cells. *Leukemia* **31**, 1000–1003 (2017).

[255] Bogani, C., Guglielmelli, P., Antonioli, E., Pancrazzi, A., Bosi, A. et al. B-, T-, and NK-cell lineage involvement in JAK2V617F-positive patients with idiopathic myelofibrosis. *Haematologica* **92**, 258–259 (2007).

[256] Larsen, T.S., Christensen, J.H., Hasselbalch, H.C. & Pallisgaard, N. The JAK2 V617F mutation involves B- and T-lymphocyte lineages in a subgroup of patients with Philadelphia-chromosome negative chronic myeloproliferative disorders. *Br J Haematol* **136**, 745–751 (2007).

[257] Buscarlet, M., Provost, S., Zada, Y.F., Bourgoin, V., Mollica, L. et al. Lineage restriction analyses in CHIP indicate myeloid bias for TET2 and multipotent stem cell origin for DNMT3A. *Blood* **132**, 277–280 (2018).

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

[259] Bolton, K.L., Ptashkin, R.N., Gao, T., Braunstein, L., Devlin, S.M. et al. Oncologic Therapy Shapes the Fitness Landscape of Clonal Hematopoiesis. *bioRxiv* 848739 (2019).

[260] Cargo, C.A., Rowbotham, N., Evans, P.A., Barrans, S.L., Bowen, D.T. et al. Targeted sequencing identifies patients with preclinical MDS at high risk of disease progression. *Blood* **126**, 2362–2365 (2015).

[261] Kwok, B., Hall, J.M., Witte, J.S., Xu, Y., Reddy, P. et al. MDS-associated somatic mutations and clonal hematopoiesis are common in idiopathic cytopenias of undetermined significance. *Blood* **126**, 2355–2361 (2015).

[262] Blackburn, E. Cancer interception. *Cancer prevention research (Philadelphia, Pa.)* **4** (2011). Publisher: Cancer Prev Res (Phila).

[263] Hait, W.N. & Levine, A.J. Genomic complexity: a call to action. *Sci Transl Med* **6**, 255cm10 (2014).

[264] Frakt, A. Why Preventing Cancer Is Not the Priority in Drug Development. *The New York Times* (2015).

[265] Amatangelo, M.D., Quek, L., Shih, A., Stein, E.M., Roshal, M. et al. Enasidenib induces acute myeloid leukemia cell differentiation to promote clinical response. *Blood* **130**, 732–741 (2017).

[266] Itzykson, R., Kosmider, O., Cluzeau, T., Mansat-De Mas, V., Dreyfus, F. et al. Impact of TET2 mutations on response rate to azacitidine in myelodysplastic syndromes and low blast count acute myeloid leukemias. *Leukemia* **25**, 1147–1152 (2011).

[267] Bejar, R., Lord, A., Stevenson, K., Bar-Natan, M., Pérez-Ladaga, A. et al. TET2 mutations predict response to hypomethylating agents in myelodysplastic syndrome patients. *Blood* **124**, 2705–2712 (2014).

[268] Cimmino, L., Dolgalev, I., Wang, Y., Yoshimi, A., Martin, G.H. et al. Restoration of TET2 Function Blocks Aberrant Self-Renewal and Leukemia Progression. *Cell* **170**, 1079–1095.e20 (2017).

[269] Steensma, D.P., Wermke, M., Klimek, V.M., Greenberg, P.L., Font, P. et al. Phase I First-in-Human Dose Escalation Study of the oral SF3B1 modulator H3B-8800 in myeloid neoplasms. *Leukemia* **35**, 3542–3550 (2021).

[270] Watts, J.M., Bradley, T.J., Thomassen, A., Brunner, A.M., Minden, M.D. et al. A Phase I/II Study to Investigate the Safety and Clinical Activity of the Protein Arginine Methyltransferase 5 Inhibitor GSK3326595 in Subjects with Myelodysplastic Syndrome and Acute Myeloid Leukemia. *Blood* **134**, 2656 (2019).

[271] Sallman, D.A., DeZern, A.E., Garcia-Manero, G., Steensma, D.P., Roboz, G.J. et al. Eprenetapopt (APR-246) and Azacitidine in TP53-Mutant Myelodysplastic Syndromes. *J Clin Oncol* **39**, 1584–1594 (2021).

[272] Lewis, E.J. PRIMA-1 as a cancer therapy restoring mutant p53: a review. *Bioscience Horizons: The International Journal of Student Research* **8**, hzv006 (2015).

[273] Ostrem, J.M., Peters, U., Sos, M.L., Wells, J.A. & Shokat, K.M. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature* **503**, 548–551 (2013).

[274] Rosenthal, R., Cadieux, E.L., Salgado, R., Bakir, M.A., Moore, D.A. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).

[275] Zapata, L., Pich, O., Serrano, L., Kondrashov, F.A., Ossowski, S. et al. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biology* **19**, 67 (2018).

[276] Cook, E.K., Izukawa, T., Young, S., Rosen, G., Jamali, M. et al. Comorbid and inflammatory characteristics of genetic subtypes of clonal hematopoiesis. *Blood Adv* **3**, 2482–2486 (2019).

[277] Bick, A.G., Weinstock, J.S., Nandakumar, S.K., Fulco, C.P., Bao, E.L. et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).

[278] Cook, E.K., Luo, M. & Rauh, M.J. Clonal hematopoiesis and inflammation: Partners in leukemogenesis and comorbidity. *Exp Hematol* **83**, 85–94 (2020).

[279] Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. (2019).

[280] Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J. et al. Ensembl 2021. *Nucleic Acids Research* **49**, D884–D891 (2021).

A

Supplementary material for Chapter 2

A.1 Theoretical predictions of a branching model of HSC dynamics

Supplementary material for Section 2.2

To determine the expected distribution of clone sizes (VAFs) at a particular time (t), under the branching process model of HSC dynamics, it is important to consider:

- How large a clone is that has survived to time t
- How likely is it that a clone survives to time t

A.1.1 Neutral mutations ($s = 0$)

We first consider the behaviour of cells that have acquired a neutral mutation, and are thus behaving according to normal homeostasis (i.e. when the average offspring per stem cell division is 1).

How large is a neutral clone that has survived to time t?

If the variance in the number of offspring from 1 cell per generation, is σ^2, then the variance in the number of offspring from n cells per generation, is $n\sigma^2$ and after t generations is $tn\sigma^2$. If, for example, we want to know how long it takes for the number of cells to double (i.e. n to increase to n^2), then:

$$n^2 = tn\sigma^2 \quad \Rightarrow \quad t = \frac{n}{\sigma^2} \quad \Rightarrow \quad t = n \tag{A.1}$$

$$\text{(for a Poisson distribution, } \sigma^2 = 1\text{)}$$

Therefore it takes approximately n generations for a clone of size n to change by n cells. From this we can infer, that if there are n cells in a clone (with no fitness effect), then the clone originated as a single cell $t = n$ generations ago and grew roughly linearly by drift.

How likely is it that a neutral clone survives to time t, due to drift?

Because the average offspring per stem cell per generation is 1, we can infer:

$$1 = P(\text{alive}|t) \times \text{size at time } t \text{ if alive}) + P(\text{extinct}|t) \times \text{size at time } t \text{ if extinct})$$

$$1 = P(\text{alive}|t) \times t) + P(\text{extinct}|t) \times 0)$$

$$P(\text{alive}|t) = \frac{1}{t}$$

$$\text{because we know } t = n, \quad P(n|t) = \frac{1}{n} \tag{A.2}$$

Therefore for a single neutral clone, the probability of still being alive after time t is $\sim \frac{1}{t}$ and the size of the clone is t.

What is the distribution of VAFs at time t, due to drift?

Clones with sizes in the interval range dn must have entered the population n generations ago, with the probability of the clone surviving being $\frac{1}{n}$. Therefore, if a neutral mutation occurs in the cells at a rate of $N\mu$ (total number of HSCs × mutation rate), the number of mutant cells alive in the interval dn can be calculated as follows:

$$\text{Probability of mutant cells being alive at time } t = \frac{1}{n}$$

$$\text{Number of mutant cells born } n \text{ generations ago } = N\mu dt$$

$$\text{Number of mutant clones alive in the interval } dn = P(\text{alive}|t) \times \text{number born}$$

$$= \frac{N\mu}{n} dn$$

This assumes that clones can be infinitely old, when in fact the oldest a clone can be after time t is t and so there is an exponential 'cut-off' due to the distribution of those single mutants that entered at time $t = 0$, i.e:

$$= \frac{N\mu}{n} e^{\frac{n}{t}} dn \tag{A.3}$$

Changing 'cell numbers' to variant allele frequencies (f) requires the substitution:

$$f = \frac{n}{2N} \qquad \text{and} \qquad df = \frac{dn}{2N} \qquad \text{and} \qquad dn = 2Ndf$$

This gives the following distribution for the VAFs of neutral mutations (changing by drift):

$$\rho(f) = \frac{2N\mu}{f} e^{-\frac{f}{\phi}} \qquad \text{where} \qquad \phi = \frac{t}{2N} \tag{A.4}$$

A.1.2 Beneficial mutations ($s > 0$)

We next consider how the behaviour of cells changes if a beneficial mutation ($s > 0$) is acquired, such that the average offspring per generation changes from 1 to $1 + s$.

How large is a beneficial clone that has survived to time t?

When considering the characteristic size of a clone with a beneficial mutation, that has been alive for t generations, we have to account for exponential growth, such that characteristic size of the clone after time t is:

$$n \sim \frac{e^{st} - 1}{s} \tag{A.5}$$

When $st \ll 1$ (i.e. early in the clone's life):

$$\frac{e^{st}-1}{s} \approx \frac{st}{s} \Rightarrow n \sim t \tag{A.6}$$

So, when $st \ll 1$), the clone behaves neutrally and grows linearly with time.

When $st \gg 1$, however, the size of the clone is dominated by:

$$n \sim \frac{e^{st}}{s} \tag{A.7}$$

Therefore when $st \gg 1$, selection dominates over drift and the clone grows exponentially over time. The point at which this occurs ($1/s$) is called the 'establishment' time (Figure A.1). The greater the fitness effect (s), the quicker clones will 'establish'.

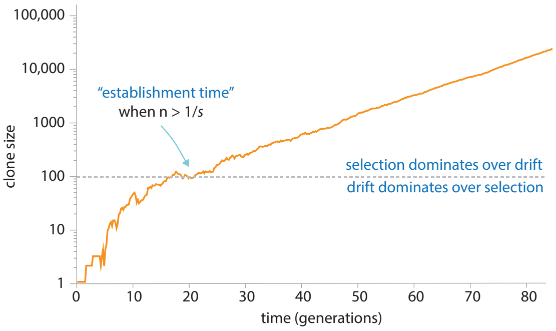


Figure A.1 Growth trajectory of a mutant clone with fitness effect s. When $n < 1/s$, drift dominates over selection and the mutant clone grows linearly with time. When $n > 1/s$, the clone 'establishes' and grows exponentially.

What is the distribution of VAFs at time t for beneficial mutations?

This distribution of VAFs for beneficial mutations is the same as for neutral mutations, except ϕ now takes into account the exponential growth of the clones:

$$\rho(f) = \frac{2N\mu}{f} e^{-\frac{f}{\phi}} \quad \text{where} \quad \phi = \frac{e^{st}-1}{2Ns} \tag{A.8}$$

This expression, for the distribution of VAFs, assumes that the expanding clone does not contribute to the total number of cells. A more accurate expression is to change $2f = \frac{n}{N}$ to $2f = \frac{n}{(n+N)}$, giving:

$$\rho(f) = \frac{2N\mu}{f(1-2f)} e^{-\frac{f}{\phi(1-2f)}} \quad \text{where} \quad \phi = \frac{e^{st}-1}{2Ns} \tag{A.9}$$

Considering time in years, rather than HSC generations, this gives the expression shown in the main text:

$$\rho(f) = \frac{2N\tau\mu}{f(1-2f)} e^{-\frac{f}{\phi(1-2f)}} \quad \text{where} \quad \phi = \frac{e^{st}-1}{2N\tau s} \tag{A.10}$$

A.2 Study panel footprints for DNMT3A, TET2, ASXL1 and TP53

Supplementary material for Section 2.3.1

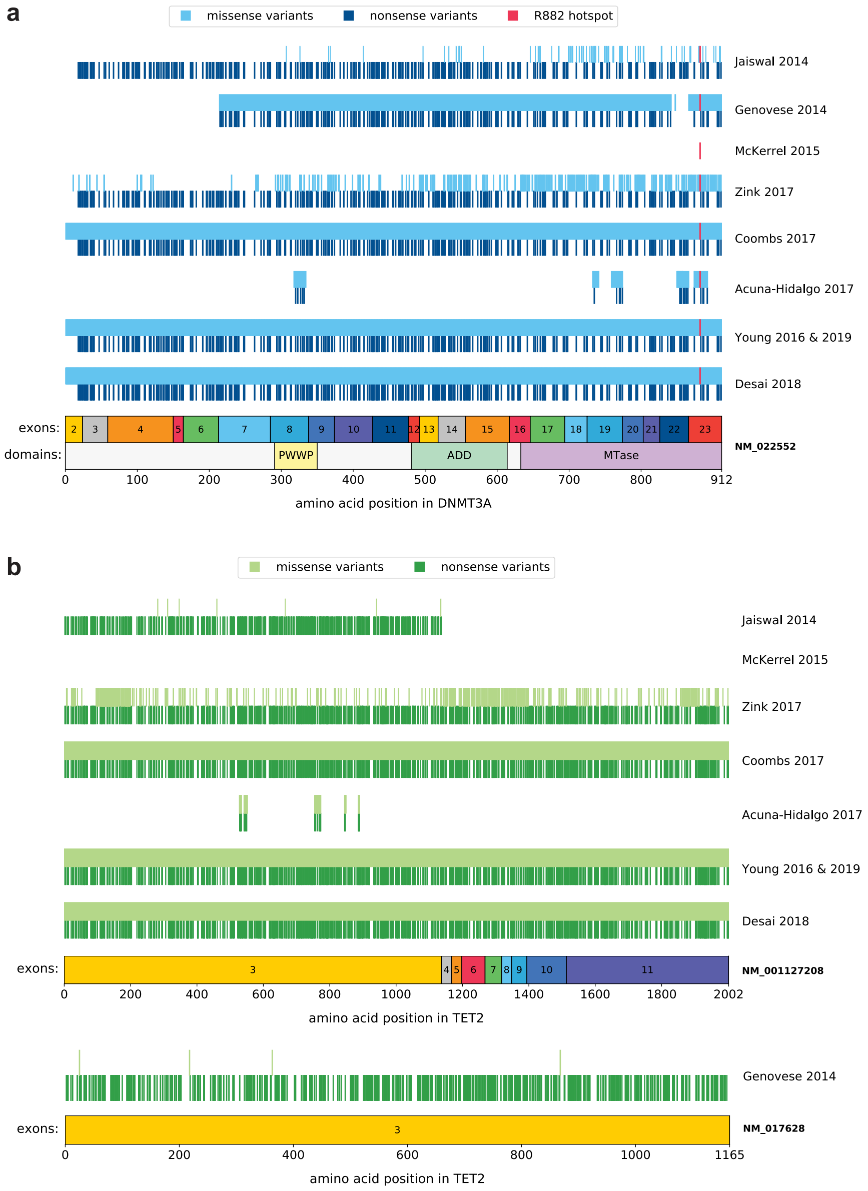


Figure A.2 Panel 'footprint' for DNMT3A and TET2 for each of the studies included in our analysis. Each vertical line represents an amino acid position at which a variant had the potential to be called in that study. A missense variant is possible at any amino acid position and so an absence of missense vertical lines represents sites that were not included by the study. Not all amino acid positions have potential nonsense variants and so an absence of nonsense vertical lines could either represent a position that does not have the potential to have a nonsense variant or a position that was not included by the study. **a. DNMT3A panel footprints.** Missense variants shown in light blue, nonsense variants in dark blue and DNMT3A R882H variants in red. PWWP: Pro-Trp-Trp-Pro domain, ADD: ATRX-DNMT3A-DNMT3L domain, MTase: Methyltransferase domain. **b. TET2 panel footprints.** Missense variants shown in light green, nonsense variants in dark green. The majority of studies annotated variants using the NM_001127208 transcript (top plot), although the NM_017628 transcript was used by Genovese 2014[4] (bottom plot). McKerrel 2015[6] did not target TET2 in their panel.

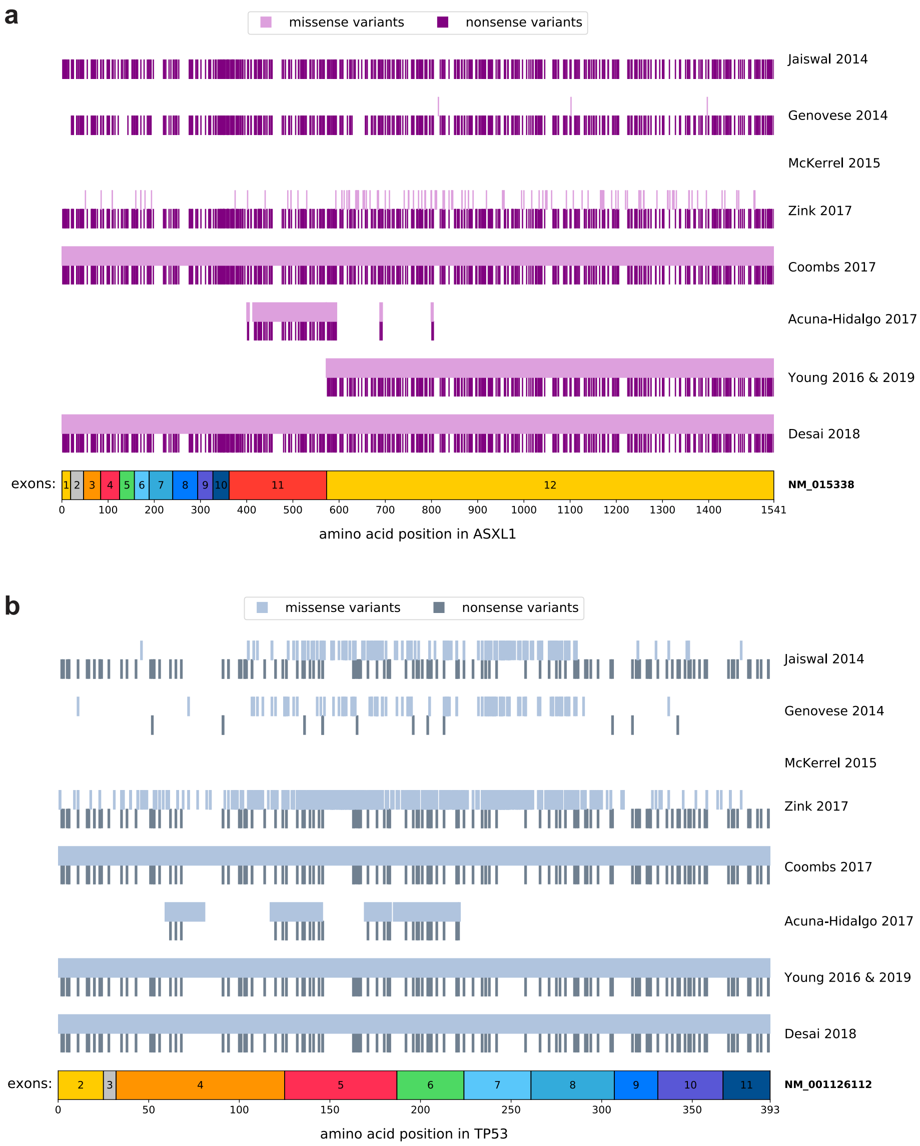


Figure A.3 Panel 'footprint' for ASXL1 and TP53 for each of the studies included in our analysis. Each vertical line represents an amino acid position at which a variant had the potential to be called in that study. A missense variant is possible at any amino acid position and so an absence of missense vertical lines represents sites that were not included by the study. Not all amino acid positions have potential nonsense variants and so an absence of nonsense vertical lines could either represent a position that does not have the potential to have a nonsense variant or a position that was not included by the study. **a. ASXL1 panel footprints.** Missense variants shown in light purple, nonsense variants in dark purple. The majority of studies annotated variants using the NM_015338 transcript. McKerrel 2015[6] did not target ASXL1 in their panel. **b. TP53 panel footprints.** Missense variants shown in light grey, nonsense variants in dark grey. The majority of studies annotated variants using the NM_001126112 transcript. McKerrel 2015[6] did not target TP53 in their panel.

A.3 Mutation rate estimates

A.3.1 Trinucleotide frequencies across the genome

Supplementary material for Section 2.3.1

Trinucleotide frequencies across the genome were calculated in R/Bioconductor using the *Biostrings* package[279], using *BS.genome.Hsapiens.UCSC.hg19*.

Table A.1 Trinculeotide frequencies across the genome.

Site	% of genome	Site	% of genome	Site	% of genome	Site	% of genome
AAA	3.8356	**CAA**	1.8905	**GAA**	1.9703	**TAA**	2.0782
AAC	1.4548	**CAC**	1.5012	**GAC**	0.9439	**TAC**	1.1341
AAG	1.9933	**CAG**	2.0254	**GAG**	1.6834	**TAG**	1.2903
AAT	2.4910	**CAT**	1.8364	**GAT**	1.3358	**TAT**	2.0626
ACA	2.0131	**CCA**	1.8426	**GCA**	1.4395	**TCA**	1.9583
ACC	1.1622	**CCC**	1.3141	**GCC**	1.1901	**TCC**	1.5432
ACG	0.2510	**CCG**	0.2761	**GCG**	0.2381	**TCG**	0.2209
ACT	1.6076	**CCT**	1.7769	**GCT**	1.3983	**TCT**	2.2134
AGA	2.2100	**CGA**	0.2205	**GGA**	1.5441	**TGA**	1.9589
AGC	1.3978	**CGC**	0.2379	**GGC**	1.1896	**TGC**	1.4408
AGG	1.7749	**CGG**	0.2761	**GGG**	1.3154	**TGG**	1.8462
AGT	1.6097	**CGT**	0.2516	**GGT**	1.1636	**TGT**	2.0205
ATA	2.0605	**CTA**	1.2886	**GTA**	1.1347	**TTA**	2.0814
ATC	1.3343	**CTC**	1.6834	**GTC**	0.9452	**TTC**	1.9728
ATG	1.8365	**CTG**	2.0269	**GTG**	1.5049	**TTG**	1.8981
ATT	2.4945	**CTT**	1.9972	**GTT**	1.4607	**TTT**	3.8502

A.3.2 Variant and study-specific mutation rates

Supplementary material for Section 2.3.1

Table A.2 Variant-specific mutation rates for the top 20 most commonly observed variants. Calculated using the site-specific mutation rates (Table 2.2 in Section 2.3.1) for the nucleotide change (and its trinucleotide context) that gives rise to the variant.

Variant	Trinucleotide Context	μ ($\times 10^{-9}$ /year)
DNMT3A R320*	C[C>T]G	14.15
DNMT3A R326C	C[C>T]G	14.15
DNMT3A R598*	G[C>T]G	18.82
DNMT3A R729W	C[C>T]G	14.15
DNMT3A Y735C	T[A>G]C	0.88
DNMT3A R736C	C[C>T]G	14.15
DNMT3A R736H	C[G>A]C	18.82
DNMT3A R771*	G[C>T]G	18.82
DNMT3A R882C	C[C>T]G	14.15
DNMT3A R882H	C[G>A]C	18.82
DNMT3A W860R	A[T>A]G, A[T>C]G	1.99
DNMT3A P904L	C[C>T]G	14.15
GNB1 K57E	C[A>G]A	0.54
IDH2 R140Q	C[G>A]G	14.15
JAK2 V617F	T[G>T]T	1.33
SF3B1 K666N	A[G>T]A, A[G>C]A	0.97
SF3B1 K700E	G[A>G]A	0.54
SRSF2 P95H	C[C>A]C	0.81
SRSF2 P95L	C[C>T]C	2.65
SRSF2 P95R	C[C>G]C	0.46

Table A.3 Study-specific mutation rates for non-synonymous DNMT3A, TET2, ASXL1, TP53 variants and all synonymous variants. Calculated by summing the site-specific mutation rates (Table 2.2 in Section 2.3.1) across the regions of the gene covered by each study.

	Non-synonymous μ ($\times 10^{-9}$ /year)				Synonymous μ
	DNMT3A	TET2	ASXL1	TP53	($\times 10^{-9}$ /year)
Jaiswal 2014[3]	894	523	625	787	-
Genovese 2014[4]	6130	557	609	609	-
McKerrel 2015[6]	36	-	-	-	-
Zink 2017[9]	1600	1960	1140	1620	-
Coombs 2017[75]	8470	13100	12000	3430	-
Acuna-Hidalgo 2017[8]	819	313	178	927	2730
Young 2016 & 2019[7,39]	8470	13000	7100	3430	85000
Desai 2018[38]	8470	13000	11900	3430	-

A.4 Fitness landscape of clonal haematopoiesis

A.4.1 Parameter estimation for the top 20 most commonly observed CH variants

Supplementary material for Section 2.5.1

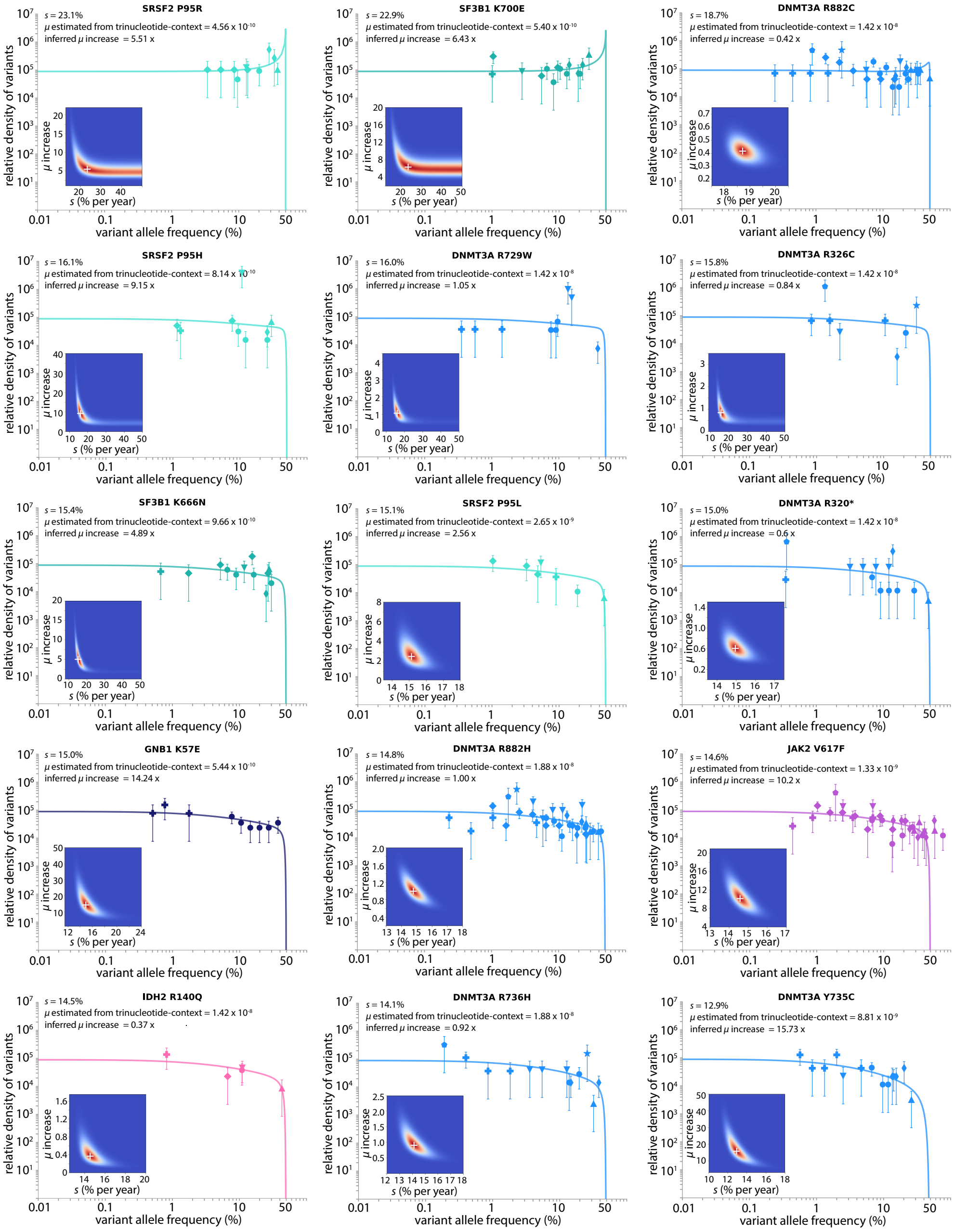


Figure A.4 Parameter estimation for the top 20 most commonly observed CH variants: part 1. $N\tau$ and the standard deviation of ages (σ) were fixed to that inferred from DNMT3A R882H and maximum likelihood approaches were used to infer s for each variant, as well as the increase of μ relative to the μ estimated from the variant's site-specific trinucleotide context (Table A.2). Each study is represented by a shaped symbol as described in Figure 2.6 (Section 2.4).

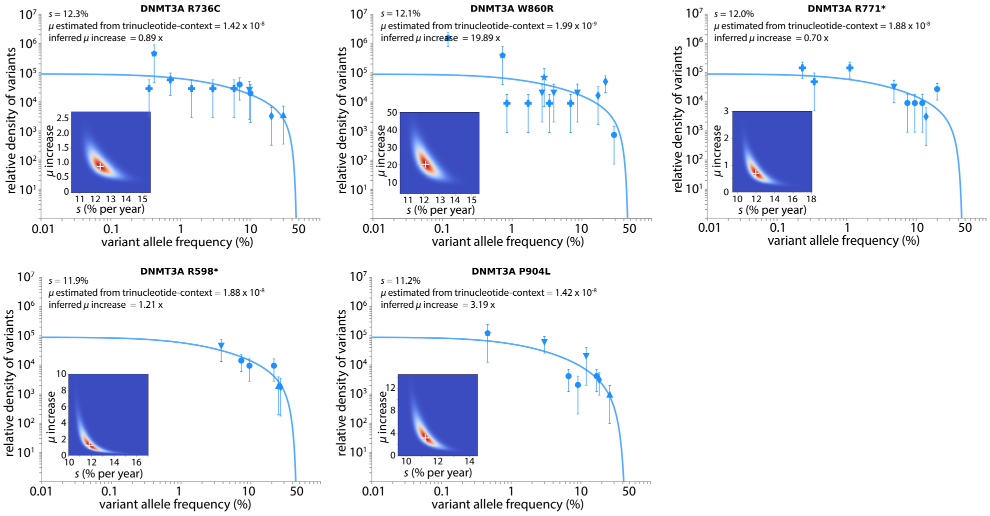


Figure A.5 Parameter estimation for the top 20 most commonly observed CH variants: part 2. $N\tau$ and the standard deviation of ages (σ) were fixed to that inferred from DNMT3A R882H and maximum likelihood approaches were used to infer s for each variant, as well as the increase of μ relative to the μ estimated from the variant's site-specific trinucleotide context (Table A.2). Each study is represented by a shaped symbol as described in Figure 2.6.

A.4.2 Parameter estimation for distribution of fitness effects in CH genes

Supplementary material for Section 2.5.2

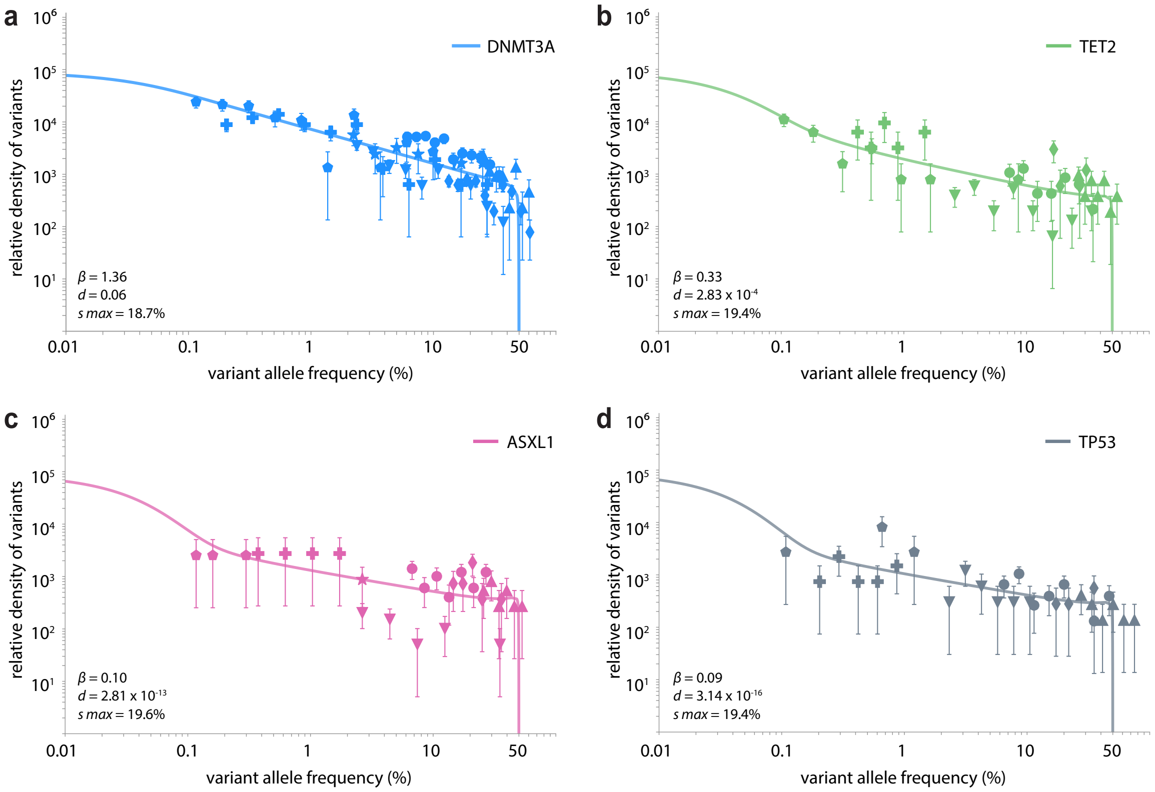


Figure A.6 Parameter estimation for distribution of fitness effects of nonsynonymous variants within commonly mutated CH genes. **a.** DNMT3A. **b.** TET2. **c.** ASXL1. **d.** TP53. Each study is represented by a shaped symbol as described in Figure 2.6 (Section 2.4).

A.5 Highly fit variants are enriched in pre-AML

Supplementary material for Section 2.12

Table A.4 Number of individuals with high-fitness or lower-fitness variants in Desai 2018[38], Abelson 2018[77] and Young 2019[39]. 'High-fitness variant' refers to any of the 20 highly fit variants we identified in Figure 2.8 (Section 2.5.1). 'Lower-fitness variant' refers to any other single nucleotide variant (SNV). If an individual had both a high-fitness and a low-fitness variant they were included in only the 'high-fitness' category.

Desai 2018[38]	**AML cases**	**controls**
no. of individuals in study	*188*	*181*
High fitness variant	63	11
Lower fitness variant	66	37

Abelson 2018[77]	**AML cases**	**controls**
no. of individuals in study	*124*	*676*
High fitness variant	27	28
Lower fitness variant	56	184

Young 2019[39]	**AML cases**	**controls**
no. of individuals in study	*34*	*69*
High fitness variant	16	15
Lower fitness variant	18	49

TOTAL	**AML cases**	**controls**
no. of individuals	*346*	*926*
High fitness variant	106	54
Lower fitness variant	148	306

A.6 Estimating fitness effects of infrequently mutated sites

Supplementary material for Section 2.8

Table A.5 Fitness effects of DNMT3A variants estimated using a crude counting method to infer the fitness effect required to achieve the actual number of observations of the variant. Variants that are within the top 20 observed variants in clonal haematopoiesis (Figure 2.8, Section 2.5.1) are highlighted in red. Range of s was calculated using the sampling noise of the number of observed variants. Site-specific mutation rates are those calculated from trinucleotide context (Table 2.2, Section 2.3.1). The number of times the variant is seen in COSMIC v87[166] (haematopoietic and lymphoid cancers) as well as their frequencies in ExAC[186] are shown.

DNMT3A variant	s (% /year)	Range of s (% /year)	Observed number	Site-specific mutation rate ($\times 10^{-9}$ /year)	Number of times in COSMIC v87	ExAC frequencies (%)						
						Overall	African	East Asian	European	Finnish	Latino	South Asian
p.E30A	28.05	20.7 - 196.75	4	0.19	3	0.2959	0.02	0	0.3033	2.1424	0.1413	0.0247
p.F751C	21.62	14.91 - 27.48	2	0.16	-	-	-	-	-	-	-	-
p.Y735C	19.93	18.54 - 21.24	23	0.88	12	0.0118	0.0097	0	0.0182	0	0.0087	0
p.F732C	18.41	13.86 - 21.62	2	0.19	1	-	-	-	-	-	-	-
p.S714C	17.63	15.68 - 19.31	7	0.44	25	0.0041	0	0	0.0075	0	0	0
p.F414C	17.23	12.95 - 20.37	2	0.16	-	-	-	-	-	-	-	-
p.L737R	15.58	11.98 - 17.91	2	0.19	1	-	-	-	-	-	-	-
p.P507R	15.56	11.66 - 18.09	2	0.46	-	-	-	-	-	-	-	-
p.W795S	15.48	11.61 - 18.00	2	0.46	-	-	-	-	-	-	-	-
p.W860R	15.45	14.46 - 16.34	18	1.99	7	-	-	-	-	-	-	-
p.Y908C	15.04	12.16 - 17.53	4	1.47	-	0.0033	0	0	0.0045	0	0.0086	0
p.F755S	14.96	13.02 - 16.41	4	0.50	3	0.0041	0	0	0.006	0	0.0086	0
p.R882P	14.92	13.09 - 16.40	6	0.63	36	0.0058	0.029	0.0116	0.003	0	0.0087	0
p.N838D	14.59	11.38 - 16.45	2	0.50	5	-	-	-	-	-	-	-
p.E599D	13.93	11.89 - 15.41	4	1.42	-	0.0034	0	0	0.0061	0	0	0
p.L547H	13.72	11.03 - 15.34	2	0.38	5	0.0017	0	0	0	0	0.0087	0.0061
p.E733G	13.62	11.10 - 15.17	3	0.81	1	-	-	-	-	-	-	-
p.G543A	13.54	10.91 - 15.11	2	0.40	5	-	-	-	-	-	-	-
p.R729G	13.51	11.48 - 14.92	3	0.66	-	0.0009	0.0097	0	0	0	0	0
p.S770W	13.28	11.19 - 14.66	3	0.57	5	-	-	-	-	-	-	-
p.P307R	13.19	10.66 - 14.66	2	0.46	2	-	-	-	-	-	-	-
p.R882H	13.07	12.75 - 13.36	105	18.82	1069	0.0545	0.0869	0.0694	0.0571	0.0758	0	0.0485
p.F732S	13.02	10.55 - 14.47	2	0.50	3	-	-	-	-	-	-	-
p.E756D	12.96	9.97 - 14.73	2	0.97	-	-	-	-	-	-	-	-

Table A.5 continued from previous page

DNMT3A variant	s (% /year)	Range of s (% /year)	Observed number	Site-specific mutation rate ($\times 10^{-9}$ /year)	Number of times in COSMIC v87	ExAC frequencies (%)						
						Overall	African	East Asian	European	Finnish	Latino	South Asian
p.I780T	12.91	10.88 - 14.18	3	1.35	7	0.0051	0.0201	0	0.0061	0	0	0
p.G543C	12.81	11.04 - 13.98	3	0.81	25	-	-	-	-	-	-	-
p.M880V	12.79	11.45 - 13.81	6	1.49	7	0.0017	0.0097	0	0.0015	0	0	0
p.W753C	12.73	9.81 - 14.44	2	1.05	-	0.0008	0	0	0.0015	0	0	0
p.E733*	12.38	9.73 - 13.82	2	0.54	-	-	-	-	-	-	-	-
p.I705T	12.32	10.95 - 13.27	4	1.35	5	-	-	-	-	-	-	-
p.R882C	12.30	11.92 - 12.65	61	14.15	460	0.0355	0.0677	0.0347	0.0391	0	0.0087	0.0364
p.L639F	11.94	9.28 - 13.47	2	1.40	-	-	-	-	-	-	-	-
p.N501S	11.91	9.34 - 13.26	2	1.35	5	0.0294	0.0098	0	0.0321	0.0457	0.0174	0.0502
p.I292T	11.91	9.34 - 13.26	2	1.35	1	-	-	-	-	-	-	-
p.V665L	11.77	9.23 - 13.09	2	1.43	2	0.0008	0	0	0.0015	0	0	0
p.M439T	11.77	9.17 - 13.27	2	1.49	-	0.0025	0	0	0.0045	0	0	0
p.V657M	11.75	10.83 - 12.46	7	3.18	-	0.0085	0	0	0.0123	0	0.0088	0.0062
p.M761V	11.66	10.16 - 12.69	4	1.49	-	0.0008	0	0	0.0015	0	0	0
p.G796V	11.61	9.53 - 12.76	2	0.95	1	-	-	-	-	-	-	-
p.C710S	11.53	9.47 - 12.66	2	0.99	5	-	-	-	-	-	-	-
p.C537*	11.29	9.28 - 12.38	2	1.12	-	-	-	-	-	-	-	-
p.F752L	11.27	9.81 - 12.18	3	1.70	5	0.0008	0	0.0116	0	0	0	0
p.S770*	11.17	8.75 - 12.41	2	0.91	2	0.0017	0	0	0.003	0	0	0
p.F732L	11.11	9.14 - 12.16	2	1.24	3	-	-	-	-	-	-	-
p.S377R	11.09	8.72 - 12.46	2	1.96	-	-	-	-	-	-	-	-
p.P453L	11.09	8.71 - 12.46	2	1.96	-	0.0017	0	0	0.003	0	0	0
p.P307L	11.05	8.71 - 12.29	2	1.96	1	-	-	-	-	-	-	-
p.G699S	10.98	9.84 - 11.76	4	2.65	2	0.0033	0.0096	0	0.0045	0	0	0
p.V296M	10.91	9.20 - 12.04	3	3.18	-	-	-	-	-	-	-	-
p.C861*	10.78	8.38 - 12.04	2	1.12	-	-	-	-	-	-	-	-
p.Y536*	10.78	8.89 - 11.79	2	1.48	-	-	-	-	-	-	-	-
p.A662T	10.71	8.46 - 12.01	2	2.32	-	0.0008	0	0	0.0015	0	0	0
p.T862I	10.70	8.94 - 12.05	4	3.10	-	-	-	-	-	-	-	-
p.R326S	10.66	8.30 - 11.88	2	1.16	-	0.0008	0	0	0	0.0151	0	0
p.G332E	10.60	7.84 - 12.16	2	1.56	-	-	-	-	-	-	-	-
p.F731L	10.55	8.71 - 11.52	2	1.70	3	0.0009	0.0097	0	0	0	0	0

Table A.5 continued from previous page

DNMT3A variant	s (% /year)	Range of s (% /year)	Observed number	Site-specific mutation rate ($\times10^{-9}$ /year)	Number of times in COSMIC v87	ExAC frequencies (%)						
						Overall	African	East Asian	European	Finnish	Latino	South Asian
p.R882S	10.41	7.70 - 12.08	2	1.16	40	-	-	-	-	-	-	-
p.Q356*	10.30	8.52 - 11.24	2	1.96	2	-	-	-	-	-	-	-
p.P904L	10.18	9.65 - 10.60	13	14.15	16	0.0091	0.0096	0	0.012	0	0.0086	0.0061
p.G298R	10.09	8.05 - 11.27	2	3.10	-	-	-	-	-	-	-	-
p.Q886*	10.08	8.45 - 11.01	3	2.32	2	-	-	-	-	-	-	-
p.W409*	10.05	9.02 - 10.72	4	4.61	3	-	-	-	-	-	-	-
p.G699D	10.05	8.80 - 10.82	3	3.45	4	-	-	-	-	-	-	-
p.Q374*	10.04	8.31 - 10.94	2	2.32	1	0.0017	0	0	0	0.0155	0.0088	0
p.R729W	10.01	9.49 - 10.43	12	14.15	14	0.0043	0	0	0.0061	0	0	0.0076
p.W305*	9.88	8.88 - 10.54	4	5.14	3	0.0008	0.0096	0	0	0	0	0
p.G685E	9.84	8.15 - 10.72	2	2.65	-	-	-	-	-	-	-	-
p.R320*	9.82	9.27 - 10.26	16	14.15	5	0.0082	0.0192	0	0.0105	0	0	0.0061
p.W297*	9.60	8.42 - 10.32	3	4.61	1	-	-	-	-	-	-	-
p.C497Y	9.54	7.91 - 10.38	2	3.21	3	-	-	-	-	-	-	-
p.W330*	9.52	8.43 - 10.25	5	5.14	3	0.0008	0	0	0.0015	0	0	0
p.R736H	9.49	8.99 - 9.90	18	18.82	20	0.0067	0.0097	0	0.0091	0	0.0087	0
p.D768E	9.49	7.03 - 10.86	2	2.42	-	-	-	-	-	-	-	-
p.R598*	9.45	8.94 - 9.84	11	18.82	9	0.0059	0.01	0	0.0061	0	0	0.0121
p.R326C	9.42	8.81 - 9.89	13	14.15	2	0.0058	0.0096	0.0116	0.006	0.0151	0	0
p.R736C	9.42	8.81 - 9.89	13	14.15	21	-	-	-	-	-	-	-
p.W893*	9.41	7.80 - 10.22	2	3.52	2	0.0016	0	0	0.003	0	0	0
p.W581*	9.41	7.80 - 10.22	2	3.52	2	0.0027	0	0	0.0016	0.0177	0	0.0063
p.W601*	9.37	8.22 - 10.06	3	5.41	1	-	-	-	-	-	-	-
p.G890D	9.32	7.75 - 10.23	3	3.45	1	0.0008	0	0	0.0015	0	0	0
p.R771*	9.27	8.74 - 9.70	16	18.82	12	-	-	-	-	-	-	-
p.R749C	9.22	8.51 - 9.72	6	12.00	10	0.0025	0	0	0.0045	0	0	0
p.L547F	9.19	7.63 - 9.98	2	4.09	1	-	-	-	-	-	-	-
p.W440*	9.02	7.49 - 9.80	2	4.61	2	0.0008	0	0	0.0015	0	0	0
p.R635W	8.98	8.09 - 9.64	5	14.15	18	0.0095	0.0353	0	0	0	0	0.0128
p.W313*	8.88	7.38 - 9.63	2	5.14	3	0.0049	0.0096	0.0116	0.006	0	0	0
p.W306*	8.80	7.31 - 9.56	2	5.41	1	0.0008	0	0	0.0015	0	0	0
p.G550R	8.69	7.95 - 9.19	5	14.81	8	-	-	-	-	-	-	-

Table A.5 continued from previous page

DNMT3A variant	s (% /year)	Range of s (% /year)	Observed number	Site-specific mutation rate ($\times 10^{-9}$ /year)	Number of times in COSMIC v87	ExAC frequencies (%)						
						Overall	African	East Asian	European	Finnish	Latino	South Asian
p.P385L	8.21	7.13 - 8.90	3	14.15	-	0.005	0	0	0.0092	0	0	0
p.S770L	8.04	6.95 - 8.79	5	12.00	6	0.005	0.0096	0	0.0015	0	0.0086	0.0182
p.W327*	8.04	5.93 - 9.01	2	4.61	2	-	-	-	-	-	-	-
p.W860*	7.85	5.75 - 8.82	2	5.14	1	0.0008	0.0096	0	0	0	0	0
p.D702N	7.80	6.46 - 8.46	2	12.00	-	-	-	-	-	-	-	-
p.R326H	7.77	6.95 - 8.34	7	18.82	2	0.0025	0.0096	0	0.003	0	0	0
p.R899C	7.67	6.30 - 8.39	2	14.15	-	0.0016	0	0	0.003	0	0	0
p.G685R	7.56	6.24 - 8.20	2	14.81	4	0.0017	0	0	0.003	0	0	0
p.R899H	7.33	5.99 - 7.99	2	18.82	3	-	-	-	-	-	-	-
p.R688H	7.33	5.99 - 8.00	2	18.82	1	0.0033	0.0096	0	0.003	0	0	0.0061
p.R659H	7.30	5.99 - 7.91	2	18.82	-	0.0017	0	0	0.0015	0	0	0.0062
p.R366H	7.30	5.99 - 7.91	2	18.82	-	-	-	-	-	-	-	-
p.G332R	7.23	6.03 - 7.98	4	14.81	2	0.0041	0	0	0.0075	0	0	0
p.R771Q	6.39	4.38 - 7.30	2	12.00	4	0.0025	0	0	0.003	0	0	0

Table A.6 Fitness effects of TET2 variants estimated using a crude counting method to infer the fitness effect required to achieve the actual number of observations of the variant. Range of s was calculated using the sampling noise of the number of observed variants. Site-specific mutation rates are those calculated from trinucleotide context (Table 2.2, Section 2.3.1). The number of times the variant is seen in COSMIC v87[166] (haematopoietic and lymphoid cancers) as well as their frequencies in ExAC[186] are shown.

TET2 variant	s (% /year)	Range of s (% /year)	Observed number	Site-specific mutation rate ($\times 10^{-9}$ /year)	Number of times in COSMIC v87	ExAC frequencies (%)						
						Overall	African	East Asian	European	Finnish	Latino	South Asian
p.L532*	15.02	11.63 - 17.16	2	0.22	-	-	-	-	-	-	-	-
p.Q916*	13.59	12.19 - 14.64	5	0.99	43	0.0008	0	0	0	0	0	0.0061
p.S588*	11.9	9.74 - 13.09	2	0.82	4	-	-	-	-	-	-	-
p.Q530*	9.61	7.38 - 10.68	2	1.96	4	0.0008	0	0	0.0015	0	0	0
p.C1135Y	9.55	7.91 - 10.42	2	3.21	3	0.0008	0	0	0	0	0.0087	0
p.Q3*	9.54	7.91 - 10.38	2	3.21	-	-	-	-	-	-	-	-
p.R1516*	7.33	5.99 - 7.99	2	18.82	26	-	-	-	-	-	-	-
p.R550*	6.82	6.01 - 7.37	7	32.69	61	0.0058	0	0.0116	0.0075	0	0.0087	0
p.R1359H	6.71	5.35 - 7.31	2	32.69	6	-	-	-	-	-	-	-

Table A.7 Fitness effects of ASXL1 variants estimated using a crude counting method to infer the fitness effect required to achieve the actual number of observations of the variant. Range of s was calculated using the sampling noise of the number of observed variants. Site-specific mutation rates are those calculated from trinucleotide context (Table 2.2, Section 2.3.1). The number of times the variant is seen in COSMIC v87[166] (haematopoietic and lymphoid cancers) as well as their frequencies in ExAC[186] are shown.

ASXL1 variant	s (% /year)	Range of s (% /year)	Observed number	Site-specific mutation rate ($\times10^{-9}$ /year)	Number of times in COSMIC v87	ExAC frequencies (%)						
						Overall	African	East Asian	European	Finnish	Latino	South Asian
p.I1220F	17.28	13.85 - 19.80	3	0.50	1	0.0099	0	0	0.0181	0	0	0
p.C856*	12.43	10.13 - 13.73	2	0.64	1	0.0008	0	0	0.0015	0	0	0
p.S392*	11.95	9.84 - 13.14	2	0.82	-	-	-	-	-	-	-	-
p.Y591*	11.07	9.34 - 12.10	3	1.43	23	0.0025	0	0	0.0046	0	0	0
p.Q588*	10.99	8.59 - 12.20	2	0.99	11	-	-	-	-	-	-	-
p.Q708*	10.30	8.52 - 11.24	2	1.96	8	-	-	-	-	-	-	-
p.Q768*	10.30	8.52 - 11.24	2	1.96	3	0.0017	0	0	0.003	0	0	0
p.Q803*	10.08	8.45 - 11.01	3	2.32	2	0.0025	0	0	0.0015	0	0.0174	0
p.W1037*	9.41	7.80 - 10.22	2	3.52	1	-	-	-	-	-	-	-
p.Q432*	8.80	6.73 - 9.77	2	3.21	-	0.0025	0	0	0.0045	0	0	0
p.W583*	8.04	5.93 - 9.01	2	4.61	3	0.0017	0	0	0.003	0	0	0
p.R404*	7.91	7.13 - 8.46	7	18.82	6	0.0058	0.0096	0.0116	0.006	0.0151	0	0
p.R693*	7.77	6.95 - 8.34	7	18.82	75	0.005	0.0098	0.0233	0.0046	0	0	0
p.R417*	6.98	5.61 - 7.73	3	14.15	8	0.0074	0	0.0116	0.009	0.0302	0	0

Table A.8 Fitness effects of TP53 variants estimated using a crude counting method to infer the fitness effect required to achieve the actual number of observations of the variant. Range of s was calculated using the sampling noise of the number of observed variants. Site-specific mutation rates are those calculated from trinucleotide context (Table 2.2, Section 2.3.1). The number of times the variant is seen in COSMIC v87[166] (haematopoietic and lymphoid cancers) as well as their frequencies in ExAC[186] are shown.

TP53 variant	s (% /year)	Range of s (% /year)	Observed number	Site-specific mutation rate ($\times10^{-9}$ /year)	Number of times in COSMIC v87	ExAC frequencies (%)						
						Overall	African	East Asian	European	Finnish	Latino	South Asian
p.N235S	16.09	13.87 - 17.80	5	1.09	6	0.0239	0	0	0.0345	0.0913	0	0
p.S241C	15.00	11.67 - 16.96	2	0.44	9	0.0008	0	0	0	0.0152	0	0
p.K132E	13.60	10.13 - 15.45	2	0.54	9	0.0008	0	0	0.0015	0	0	0
p.Y220C	12.67	10.73 - 13.94	4	1.47	145	0.0025	0	0	0.0045	0	0	0
p.R110L	12.48	9.77 - 13.91	2	1.06	7	0.0017	0	0	0.0015	0	0	0.0061
p.G245C	12.27	9.61 - 13.66	2	1.16	10	-	-	-	-	-	-	-
p.V216M	8.85	6.54 - 10.10	2	3.21	60	-	-	-	-	-	-	-
p.R248Q	8.63	7.67 - 9.30	4	14.15	189	0.0058	0	0.0231	0.0075	0	0	0
p.W146*	8.04	5.93 - 9.01	2	4.61	9	-	-	-	-	-	-	-
p.R273H	7.77	7.07 - 8.27	5	32.69	106	0.0026	0	0	0.0047	0	0	0
p.R110C	7.67	6.30 - 8.39	2	14.15	-	0.0008	0	0	0	0.0151	0	0
p.R248W	7.67	6.30 - 8.39	2	14.15	129	0.0008	0	0	0.0015	0	0	0
p.R283H	7.33	5.99 - 7.99	2	18.82	2	0.0066	0	0	0.003	0	0.0262	0.0183

B

Supplementary material for Chapter 3

B.1 Distribution of mCAs across the genome

Supplementary material for Section 3.2

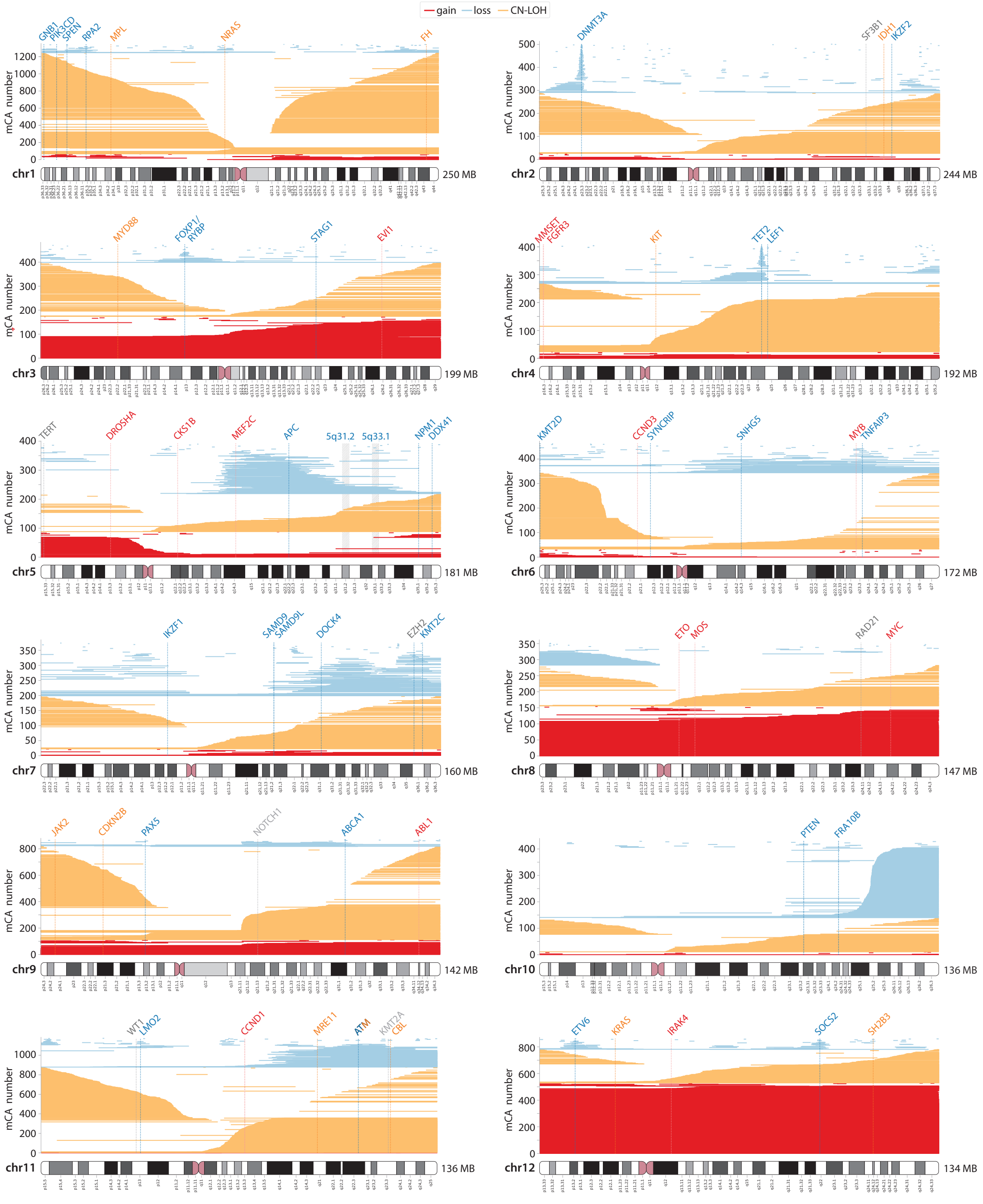


Figure B.1 mCAs detected among ∼ 500,000 UK Biobank participants: part 1. Each mCA is represented as a horizontal line. Gain events are shown in red, loss events in blue and CN-LOH events in yellow. Genes recurrently mutated in clonal haematopoiesis or haematological malignancies which may be putative target genes for loss, gain or CN-LOH events are labelled in blue, red and orange respectively.

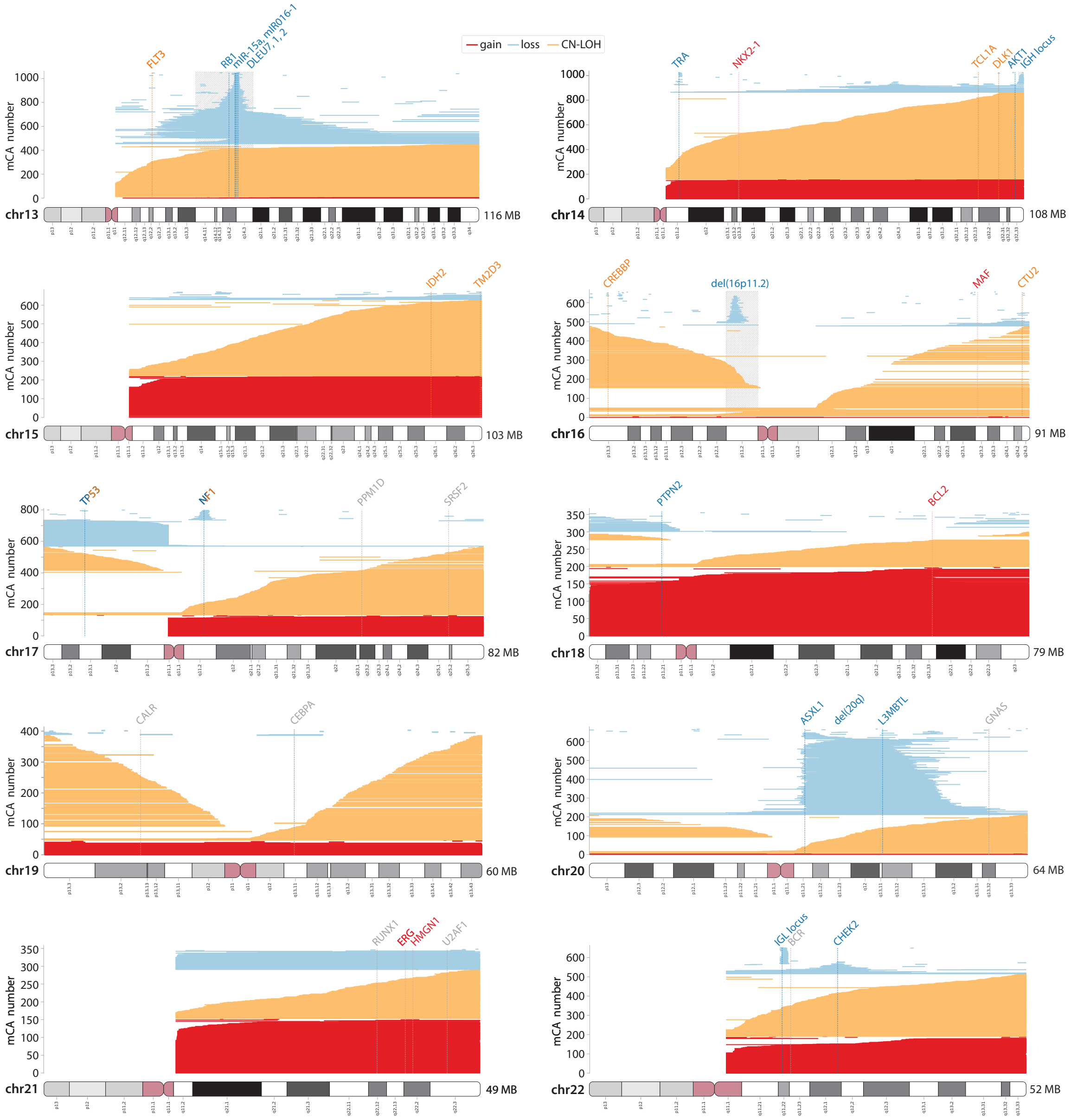


Figure B.2 mCAs detected among ∼ 500,000 UK Biobank participants: part 2. Each mCA is represented as a horizontal line. Gain events are shown in red, loss events in blue and CN-LOH events in yellow. Genes recurrently mutated in clonal haematopoiesis or haematological malignancies which may be putative target genes for loss, gain or CN-LOH events are labelled in blue, red and orange respectively.

B.2 Fitness effects and mutation rates of mCAs

Supplementary material for Section 3.4

Table B.1 Fitness effects and mutation rates for gain events. The fitness effects and mutation rates were only calculated if the mCA was observed at least 8 times. Fitness effects and mutation rates were only calculated using data from individuals who had a single mCA.

	Observed number			Fitness effect (s) (% per year)		Mutation rate (μ) ($\times 10^{-9}$/ year)	
mCA	**Single mCA**	**+ 1 other**	**+ $\geq$ 2 other**	s	s **95% C.I.**	μ	μ **95% C.I.**
1p+	23	2	1	14.19	13.13 - 17.21	0.37	0.11 - 0.70
1q+	15	7	14	14.13	12.71 - 20.43	0.15	0.03 - 0.28
2p+	8	2	6	14.74	13.27 - 47.55	0.15	0.00 - 0.37
3+	30	39	32	15.95	14.88 - 18.55	0.16	0.08 - 0.23
3p+	8	0	4	13.67	12.45 - 47.55	0.32	0.00 - 0.97
3q+	17	17	26	14.3	12.77 - 20.04	0.14	0.04 - 0.25
5+	21	0	6	9.13	8.30 - 11.35	1.66	0.34 - 3.97
5p+	32	5	4	10.3	9.52 - 11.66	1.91	0.63 - 3.66
5q+	9	5	5	15.71	13.86 - 47.64	0.08	0.01 - 0.18
6p+	13	4	5	14.47	12.59 - 46.02	0.29	0.01 - 0.74
6q+	8	0	0	12.86	11.98 - 47.52	0.72	0.01 - 2.00
8+	75	15	30	17.84	17.14 - 18.96	0.32	0.24 - 0.40
9+	46	14	10	18.44	17.43 - 20.29	0.16	0.11 - 0.22
9p+	8	5	2	13.35	11.89 - 47.46	0.11	0.00 - 0.25
9q+	18	5	4	14	12.59 - 18.96	0.15	0.04 - 0.25
12+	276	112	100	16.68	16.32 - 17.11	1.14	1.00 - 1.28
12q+	16	7	7	14.71	13.21 - 23.5	0.15	0.03 - 0.28
14q+	147	8	7	14.35	13.89 - 14.87	1.38	1.08 - 1.66
15q+	206	15	2	12.62	12.27 - 12.98	2.71	2.19 - 3.22
17q+	9	5	5	15.06	13.27 - 46.73	0.14	0.01 - 0.31
18+	47	44	80	13.84	13.04 - 15.15	0.38	0.23 - 0.52
18q+	10	7	10	15.71	13.55 - 46.12	0.07	0.01 - 0.13
21q+	125	13	14	11.15	10.73 - 11.65	2.61	1.86 - 3.34
22q+	155	23	13	11.1	10.77 - 11.48	5.17	3.72 - 6.68

Table B.2 Fitness effects and mutation rates for loss events. The fitness effects and mutation rates were only calculated if the mCA was observed at least 8 times. Fitness effects and mutation rates were only calculated using data from individuals who had a single mCA.

Observed number				Fitness effect (s) (% per year)		Mutation rate (μ) ($\times 10^{-9}$/ year)	
mCA	**Single mCA**	**+ 1 other**	**+ $\geq$ 2 other**	s	s **95% C.I.**	μ	μ **95% C.I.**
1p-	17	6	19	14.32	12.92 - 47.58	0.23	0.04 - 0.39
1q-	19	8	15	16.02	14.33 - 48.45	0.32	0.08 - 0.51
2p-	106	25	16	14.30	13.62 - 15.56	3.08	1.70 - 4.34
2q-	34	7	11	18.94	16.48 - 48.51	0.31	0.15 - 0.47
3p-	26	7	9	23.27	17.86 - 48.47	0.17	0.11 - 0.29
3q-	15	2	5	15.55	14.64 - 48.43	0.27	0.06 - 0.42
4q-	85	16	13	16.29	14.99 - 41.81	1.21	0.43 - 1.80
5q-	121	18	16	14.19	13.64 - 15.11	1.44	1.01 - 1.83
6p-	18	5	5	16.77	15.10 - 48.45	0.17	0.06 - 0.27
6q-	33	14	34	13.61	12.71 - 16.70	0.47	0.18 - 0.72
7p-	24	9	4	16.49	15.10 - 48.45	0.21	0.08 - 0.31
7q-	65	32	32	14.47	13.69 - 16.35	0.78	0.42 - 1.09
8p-	20	9	22	16.97	15.71 - 48.37	0.13	0.05 - 0.18
8q-	8	4	4	20.30	14.98 - 48.41	0.09	0.04 - 0.26
9q-	28	4	6	15.37	14.33 - 47.67	0.34	0.09 - 0.51
10q-	252	8	19	11.97	11.69 - 12.29	4.35	3.54 - 5.06
11p-	28	7	7	14.19	13.17 - 27.40	0.60	0.11 - 1.06
11q-	178	34	26	13.03	12.65 - 13.49	2.56	1.96 - 3.05
12p-	17	4	9	12.70	11.63 - 45.92	0.70	0.05 - 1.67
12q-	24	5	15	13.97	13.27 - 45.92	0.51	0.08 - 0.81
13q-	337	128	102	15.85	15.19 - 16.85	3.79	2.90 - 4.51
14q-	68	50	39	15.96	14.90 - 39.39	0.78	0.28 - 1.10
15q-	16	11	14	12.77	11.63 - 46.73	0.34	0.04 - 0.72
16p-	104	19	11	14.86	14.18 - 16.79	3.19	1.50 - 4.80
16q-	28	7	6	18.16	16.53 - 48.37	0.23	0.11 - 0.35
17p-	9	79	78	19.71	15.71 - 48.37	0.05	0.02 - 0.09
17q-	44	7	6	14.30	13.52 - 21.60	2.10	0.38 - 3.77
18p-	10	10	10	11.47	10.57 - 47.43	0.26	0.02 - 0.54
18q-	10	8	4	16.78	14.90 - 48.37	0.14	0.04 - 0.24
20-	14	0	2	9.02	8.28 - 42.01	1.51	0.03 - 3.69
20q-	364	32	24	14.21	13.85 - 14.63	6.01	4.83 - 7.06
21q-	22	4	32	13.56	12.45 - 45.10	0.22	0.05 - 0.37
22q-	60	42	33	16.40	14.90 - 46.73	0.73	0.26 - 1.06

.

Table B.3 Fitness effects and mutation rates for CNLOH events. The fitness effects and mutation rates were only calculated if the mCA was observed at least 8 times. Fitness effects and mutation rates were only calculated using data from individuals who had a single mCA.

Observed number				Fitness effect (s) (% per year)		Mutation rate (μ) ($\times 10^{-9}$/ year)	
mCA	**Single mCA**	**+ 1 other**	**+ $\geq$ 2 other**	s	**s 95% C.I.**	μ	**μ 95% C.I.**
1=	64	6	1	11.52	10.90 - 12.53	0.66	0.42 - 0.86
1p=	588	35	19	12.39	12.15 - 12.64	5.71	5.09 - 6.29
1q=	432	22	10	11.49	11.24 - 11.73	5.25	4.56 - 5.91
2=	12	1	0	7.84	6.84 - 45.41	1.01	0.02 - 3.10
2p=	95	4	2	11.99	11.50 - 12.71	1.12	0.80 - 1.43
2q=	139	8	3	11.12	10.71 - 11.63	2.21	1.64 - 2.77
3p=	106	8	4	10.85	10.42 - 11.40	1.74	1.24 - 2.26
3q=	92	7	3	10.81	10.29 - 11.43	1.75	1.20 - 2.33
4=	19	7	1	12.61	11.15 - 45.87	0.17	0.04 - 0.28
4p=	39	1	3	13.22	12.28 - 17.01	0.40	0.16 - 0.57
4q=	161	13	1	15.18	14.30 - 16.95	0.94	0.68 - 1.13
5p=	14	0	0	11.70	10.19 - 47.46	0.19	0.03 - 0.36
5q=	109	3	2	12.48	11.93 - 13.21	1.05	0.77 - 1.31
6p=	211	15	6	11.58	11.27 - 11.98	3.02	2.43 - 3.53
6q=	55	4	10	11.31	10.62 - 12.36	1.01	0.60 - 1.41
7p=	59	1	2	13.65	12.71 - 16.01	0.45	0.26 - 0.59
7q=	95	10	4	11.68	11.14 - 12.37	1.34	0.95 - 1.72
8p=	31	1	0	13.77	12.45 - 45.10	0.26	0.08 - 0.38
8q=	84	4	2	12.08	11.56 - 12.87	1.22	0.80 - 1.62
9=	59	4	5	9.44	8.82 - 10.29	1.31	0.77 - 1.89
9p=	275	38	14	14.82	14.22 - 15.76	2.28	1.75 - 2.68
9q=	286	17	7	12.86	12.48 - 13.28	2.70	2.25 - 3.05
10p=	37	2	1	13.17	12.28 - 17.31	0.30	0.13 - 0.42
10q=	74	9	4	12.17	11.56 - 13.11	0.96	0.63 - 1.25
11=	15	2	0	10.05	8.76 - 42.98	0.28	0.03 - 0.55
11p=	452	19	4	13.41	13.07 - 13.80	3.98	3.46 - 4.44
11q=	346	28	9	12.52	12.21 - 12.88	3.59	3.06 - 4.10
12=	9	1	2	6.39	5.00 - 47.00	2.47	0.02 - 17.89
12p=	35	5	1	12.04	11.14 - 14.08	0.50	0.22 - 0.77
12q=	186	8	4	12.43	12.04 - 12.96	1.80	1.44 - 2.13
13q=	380	43	20	13.04	12.69 - 13.42	3.41	2.93 - 3.83
14q=	636	44	24	12.43	12.19 - 12.66	5.98	5.31 - 6.56
15q=	383	20	4	11.15	10.89 - 11.39	5.39	4.59 - 6.01
16=	40	1	2	9.90	9.22 - 11.060	0.99	0.50 - 1.60
16p=	222	13	10	12.09	11.73 - 12.46	2.95	2.35 - 3.42
16q=	171	5	6	11.42	11.08 - 11.86	2.41	1.87 - 2.86
17=	10	1	3	8.77	8.00 - 47.00	0.39	0.02 - 0.84
17p=	84	8	6	13.16	12.52 - 14.46	0.91	0.58 - 1.17
17q=	305	13	5	12.36	12.07 - 12.73	3.28	2.75 - 3.78
18p=	14	0	0	10.05	8.76 - 44.73	0.62	0.03 - 1.78
18q=	70	6	2	11.96	11.38 - 12.85	1.05	0.66 - 1.43
19p=	139	2	6	11.36	10.98 - 11.84	2.45	1.81 - 3.05
19q=	159	18	9	12.33	11.92 - 12.84	2.14	1.61 - 2.59
20=	10	2	0	13.47	12.35 - 48.33	0.07	0.02 - 0.10
20p=	38	1	0	11.75	10.96 - 13.41	0.75	0.35 - 1.14
20q=	143	6	4	12.34	11.92 - 12.96	1.68	1.27 - 2.06
21q=	131	6	1	11.61	11.22 - 12.14	2.24	1.64 - 2.77
22q=	292	26	7	14.22	13.73 - 14.92	2.30	1.87 - 2.69

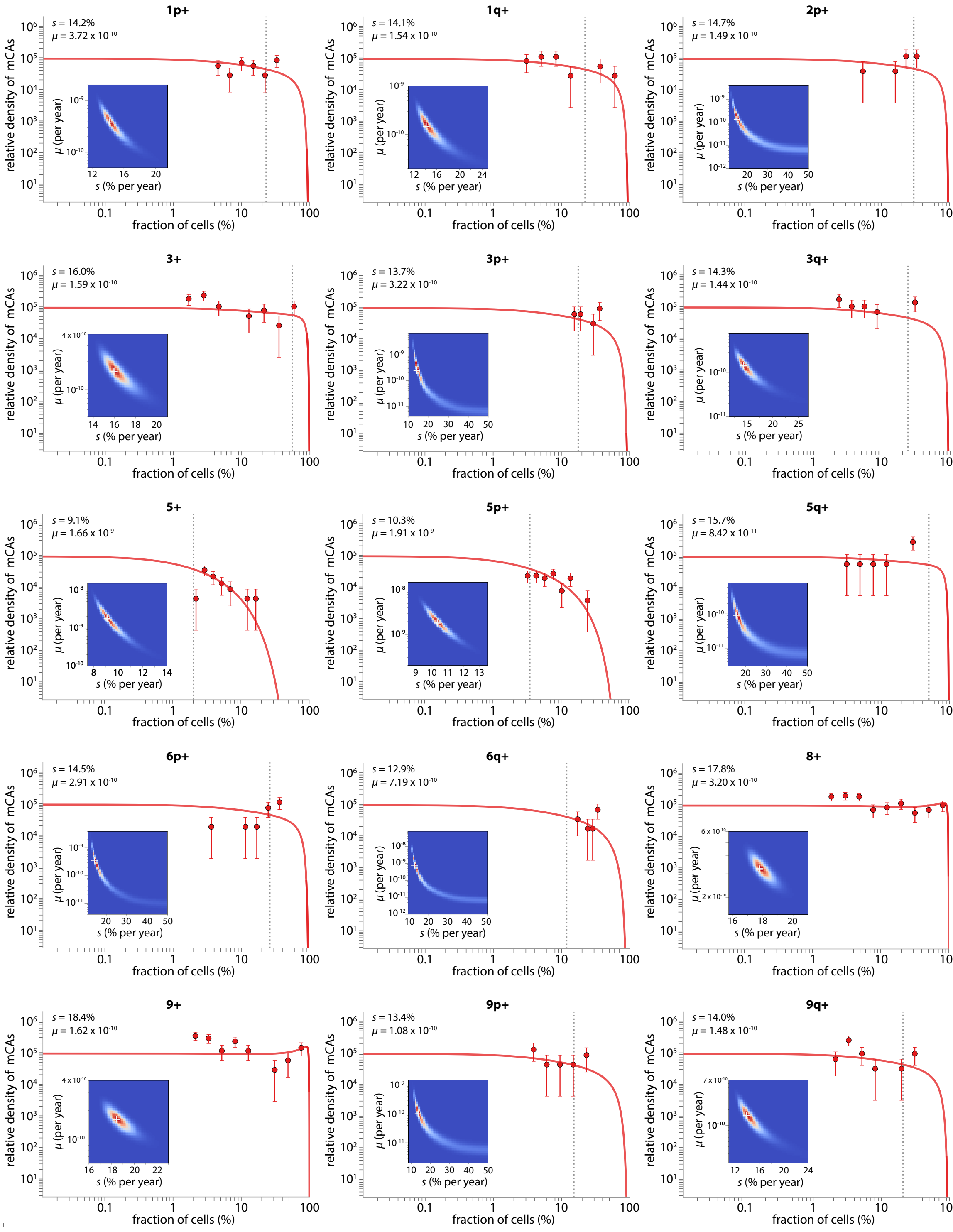


Figure B.3 Parameter estimation for individual mCAs: gains: part 1. The cell fraction probability density histogram is shown for each mCA (datapoints) with the theory distribution (solid line) fitted using maximum likelihood approaches. Error bars represent sampling noise. Grey vertical dashed line shows the fitted ϕ parameter ($\frac{e^{st}-1}{Ns}$), where the exponential fall-off in densities occurs. The white cross on the maximum likelihood heatmap marks the most likely μ and s.

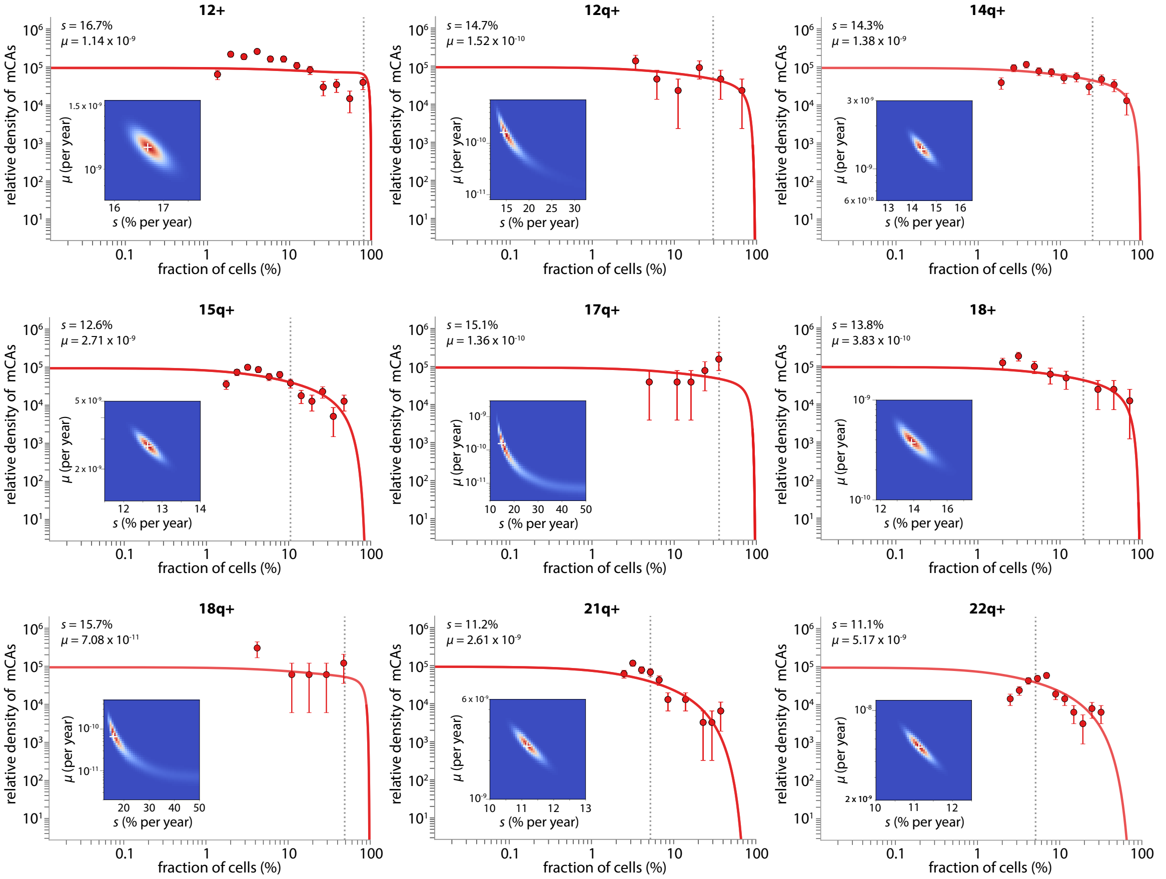


Figure B.4 Parameter estimation for individual mCAs: gains: part 1. The cell fraction probability density histogram is shown for each mCA (datapoints) with the theory distribution (solid line) fitted using maximum likelihood approaches. Error bars represent sampling noise. Grey vertical dashed line shows the fitted ϕ parameter ($\frac{e^{st}-1}{Ns}$), where the exponential fall-off in densities occurs. The white cross on the maximum likelihood heatmap marks the most likely μ and s.

.

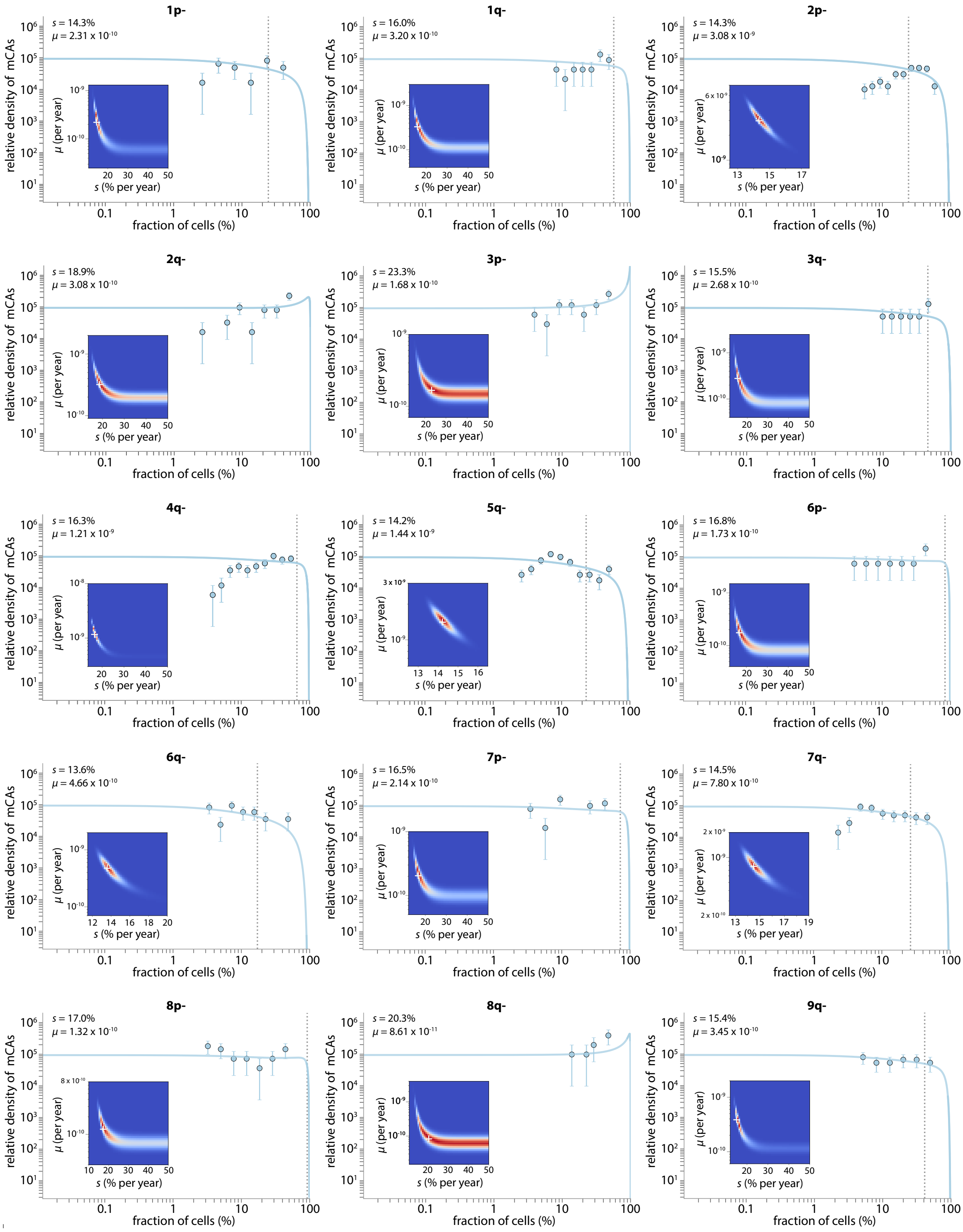


Figure B.5 Parameter estimation for individual mCAs: losses: part 1. The cell fraction probability density histogram is shown for each mCA (datapoints) with the theory distribution (solid line) fitted using maximum likelihood approaches. Error bars represent sampling noise. Grey vertical dashed line shows the fitted ϕ parameter ($\frac{e^{st}-1}{Ns}$), where the exponential fall-off in densities occurs. The white cross on the maximum likelihood heatmap marks the most likely μ and s.

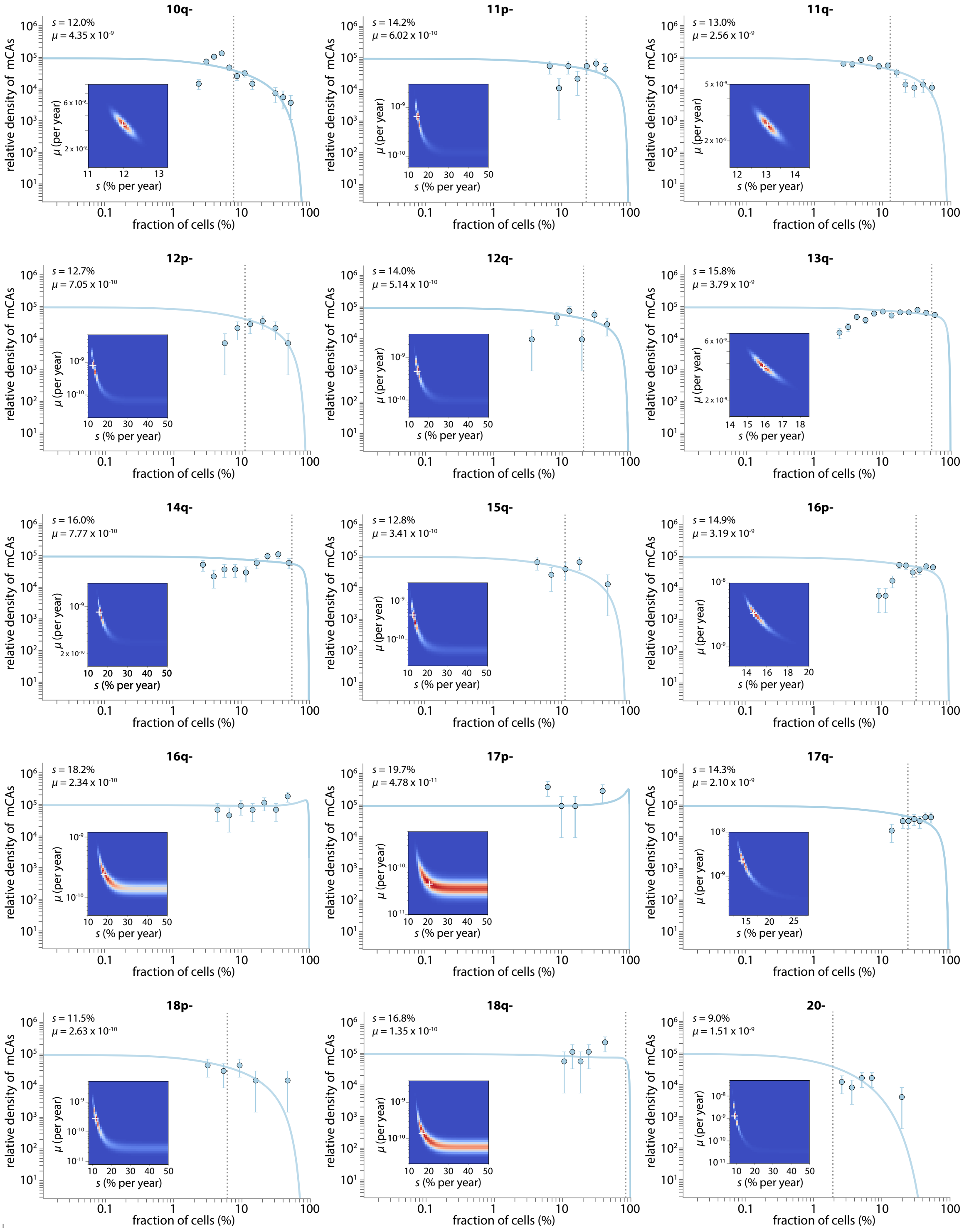


Figure B.6 Parameter estimation for individual mCAs: losses: part 2. The cell fraction probability density histogram is shown for each mCA (datapoints) with the theory distribution (solid line) fitted using maximum likelihood approaches. Error bars represent sampling noise. Grey vertical dashed line shows the fitted ϕ parameter ($\frac{e^{st}-1}{Ns}$), where the exponential fall-off in densities occurs. The white cross on the maximum likelihood heatmap marks the most likely μ and s.

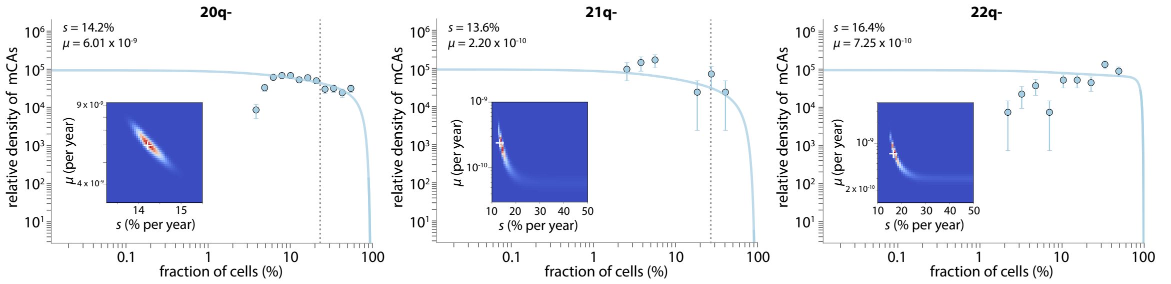


Figure B.7 Parameter estimation for individual mCAs: losses: part 3. The cell fraction probability density histogram is shown for each mCA (datapoints) with the theory distribution (solid line) fitted using maximum likelihood approaches. Error bars represent sampling noise. Grey vertical dashed line shows the fitted ϕ parameter ($\frac{e^{st}-1}{Ns}$), where the exponential fall-off in densities occurs. The white cross on the maximum likelihood heatmap marks the most likely μ and s.

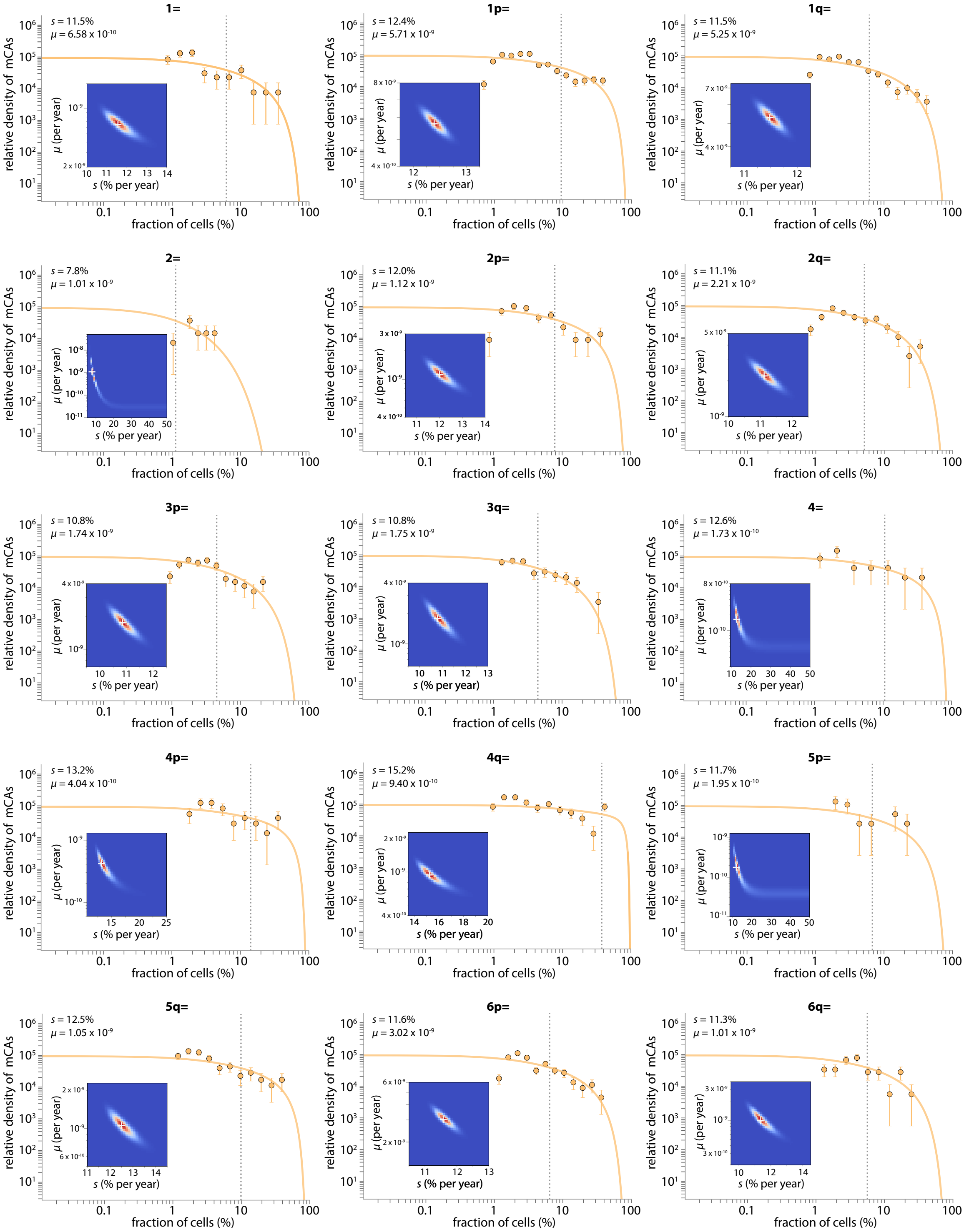


Figure B.8 Parameter estimation for individual mCAs: CN-LOH: part 1. The cell fraction probability density histogram is shown for each mCA (datapoints) with the theory distribution (solid line) fitted using maximum likelihood approaches. Error bars represent sampling noise. Grey vertical dashed line shows the fitted ϕ parameter ($\frac{e^{st}-1}{Ns}$), where the exponential fall-off in densities occurs. The white cross on the maximum likelihood heatmap marks the most likely μ and s.

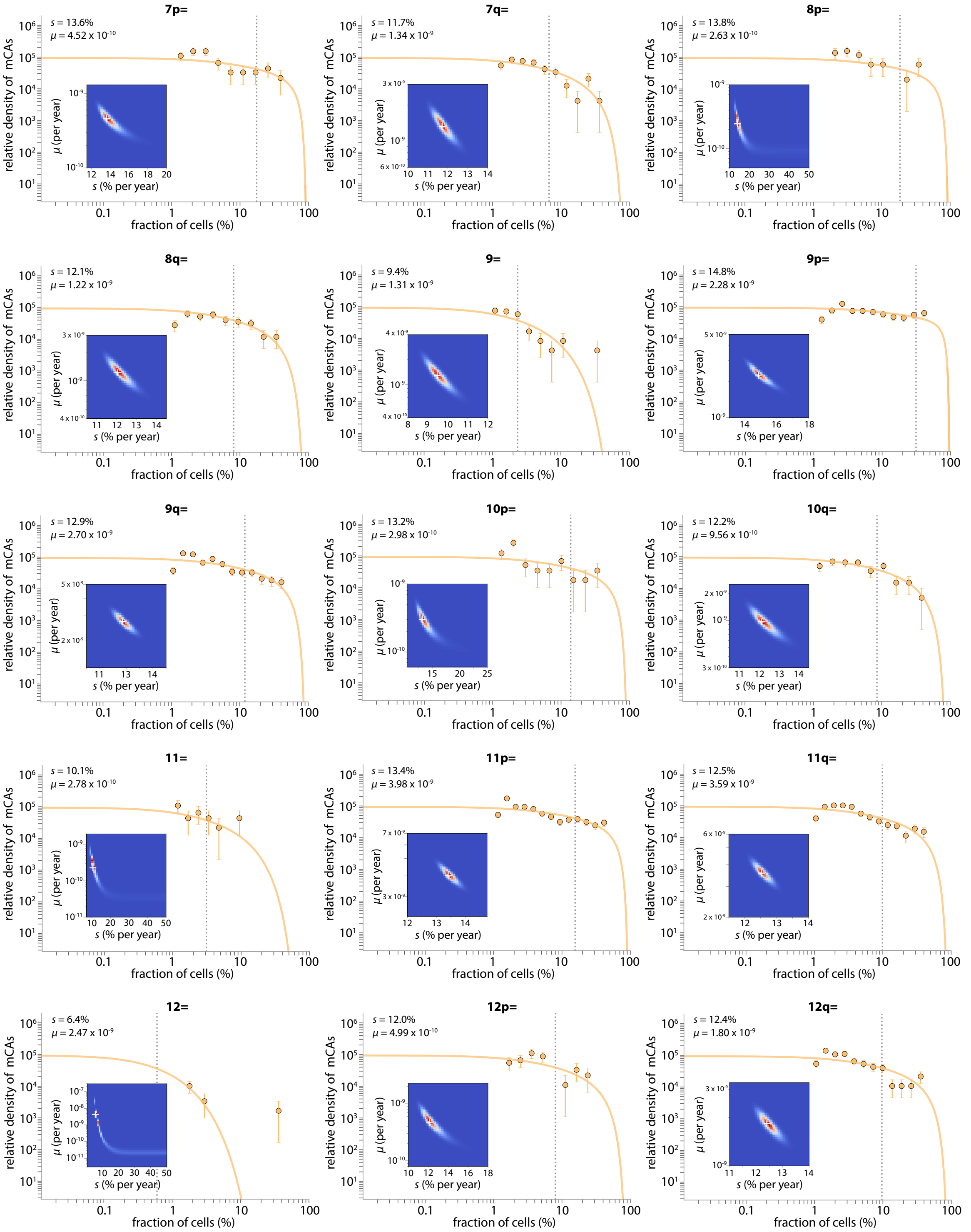


Figure B.9 Parameter estimation for individual mCAs: CN-LOH: part 2. The cell fraction probability density histogram is shown for each mCA (datapoints) with the theory distribution (solid line) fitted using maximum likelihood approaches. Error bars represent sampling noise. Grey vertical dashed line shows the fitted ϕ parameter ($\frac{e^{st}-1}{Ns}$), where the exponential fall-off in densities occurs. The white cross on the maximum likelihood heatmap marks the most likely μ and s.

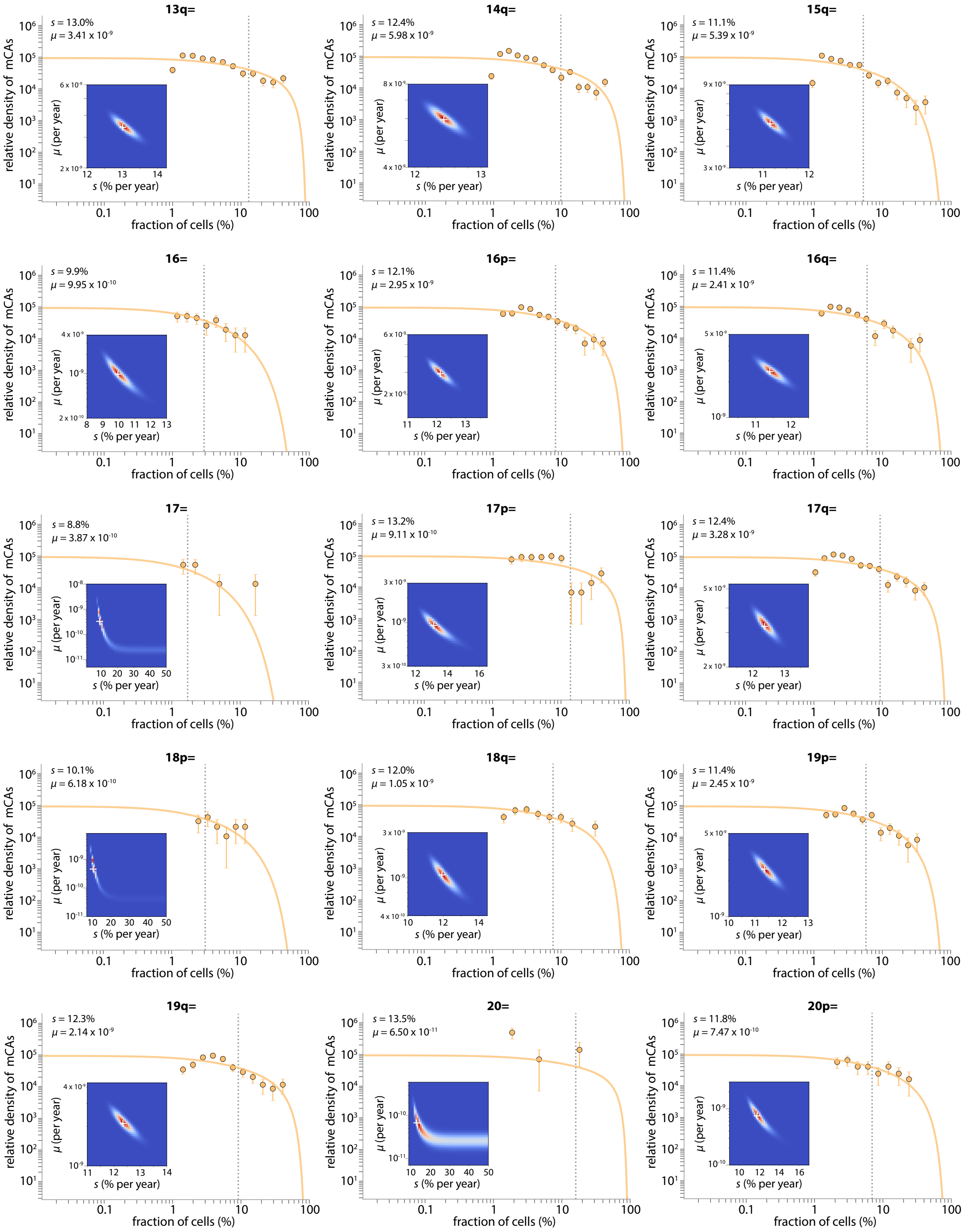


Figure B.10 Parameter estimation for individual mCAs: CN-LOH: part 3. The cell fraction probability density histogram is shown for each mCA (datapoints) with the theory distribution (solid line) fitted using maximum likelihood approaches. Error bars represent sampling noise. Grey vertical dashed line shows the fitted ϕ parameter ($\frac{e^{st}-1}{Ns}$), where the exponential fall-off in densities occurs. The white cross on the maximum likelihood heatmap marks the most likely μ and s.

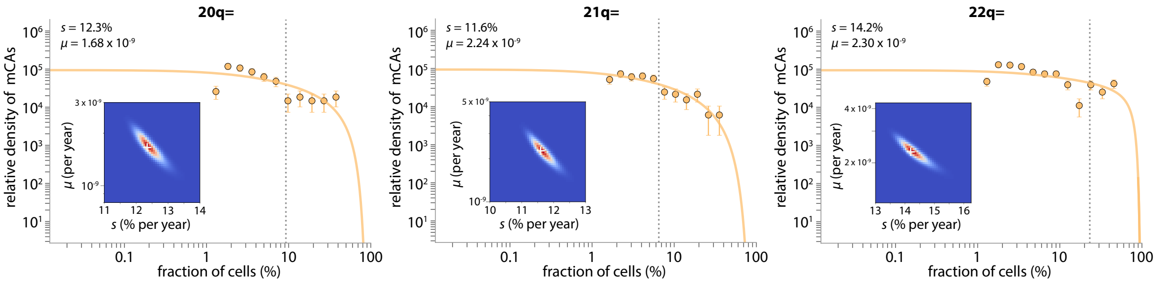


Figure B.11 Parameter estimation for individual mCAs: CN-LOH: part 4. The cell fraction probability density histogram is shown for each mCA (datapoints) with the theory distribution (solid line) fitted using maximum likelihood approaches. Error bars represent sampling noise. Grey vertical dashed line shows the fitted ϕ parameter ($\frac{e^{st}-1}{Ns}$), where the exponential fall-off in densities occurs. The white cross on the maximum likelihood heatmap marks the most likely μ and s.

B.3 Sex differences in mCA fitness effects and mutation rates

Supplementary material for Section 3.4.1

Table B.4 Sex-specific fitness effects and mutation rates for gain events. The fitness effects and mutation rates were only calculated if the mCA was observed at least 10 times. Fitness effects and mutation rates were only calculated using data from individuals who had a single mCA. The ‘observed number’ refers to the number of individuals who had the mCA as their only mCA. p-values were calculated from the area under the distribution of difference probability curve where the difference ≤ 0.

	Observed number		**Fitness effect (s) (% per year)**					**mCA-specific mutation rate (μ) ($\times10^{-9}$/ year)**				
mCA	**Men**	**Women**	**Male s**	**Male s 95% C.I.**	**Female s**	**Female s 95% C.I.**	**p-value (s)**	**Male μ**	**Male μ 95% C.I.**	**Female μ**	**Female μ 95% C.I.**	**p-value (μ)**
1p+	9	14	-	-	14.17	13.27 - 31.22	-	-	-	0.47	0.02 - 1.03	-
1q+	5	10	-	-	14.96	13.27 - 45.92	-	-	-	0.13	0.01 - 0.24	-
3+	20	10	16.14	14.85 - 20.57	7.93	6.90 - 45.10	2.3×10^{-1}	0.25	0.09 - 0.39	1.54	0.01 - 5.08	1.1×10^{-1}
5+	14	7	7.68	6.63 - 14.39	-	-	-	7.46	0.15 - 29.65	-	-	-
5p+	20	12	9.91	9.14 - 12.57	11.39	10.19 - 39.84	1.0×10^{-1}	3.58	0.56 - 9.62	0.55	0.01 - 1.23	2.5×10^{-2}
8+	30	45	17.49	16.24 - 20.29	17.27	16.29 - 18.90	3.8×10^{-1}	0.28	0.15 - 0.40	0.38	0.24 - 0.52	1.4×10^{-1}
9+	27	19	16.66	15.45 - 19.69	19.99	18.57 - 29.29	2.9×10^{-2}	0.27	0.13 - 0.38	0.09	0.03 - 0.14	5.2×10^{-3}
9q+	7	11	-	-	12.51	11.43 - 43.14	-	-	-	0.23	0.01 - 0.43	-
12+	148	128	17.17	16.64 - 17.79	14.08	13.61 - 14.65	$< 10^{-10}$	1.17	0.98 - 1.35	1.85	1.44 - 2.23	1.1×10^{-3}
12q+	3	13	-	-	15.18	13.27 - 41.02	-	-	-	0.18	0.02 - 0.36	-
14q+	87	60	14.89	14.29 - 15.66	13.15	12.46 - 14.14	3.2×10^{-3}	1.54	1.12 - 1.96	1.36	0.87 - 1.91	3.3×10^{-1}
15q+	162	44	11.75	11.42 - 12.15	14.26	13.42 - 15.77	4.7×10^{-8}	6.64	5.14 - 8.30	0.57	0.34 - 0.77	$< 10^{-10}$
18+	23	24	12.77	11.69 - 15.86	14.01	12.79 - 16.86	2.0×10^{-1}	0.49	0.18 - 0.81	0.35	0.14 - 0.55	2.3×10^{-1}
21q+	85	40	11.62	11.14 - 12.36	8.12	7.51 - 9.08	6.7×10^{-6}	3.03	2.04 - 4.01	16.33	4.94 - 35.56	1.1×10^{-3}
22q+	68	87	10.31	9.85 - 10.97	11.42	10.93 - 12.06	8.2×10^{-3}	7.51	4.24 - 11.38	4.45	2.88 - 6.21	5.5×10^{-2}

Table B.5 Sex-specific fitness effects and mutation rates for loss events. The fitness effects and mutation rates were only calculated if the mCA was observed at least 10 times. Fitness effects and mutation rates were only calculated using data from individuals who had a single mCA. The ‘observed number’ refers to the number of individuals who had the mCA as their only mCA. p-values were calculated from the area under the distribution of difference probability curve where the difference ≤ 0.

	Observed number		**Fitness effect (s) (% per year)**					**mCA-specific mutation rate (μ) ($\times10^{-9}$/ year)**				
mCA	**Men**	**Women**	**Male s**	**Male s 95% C.I.**	**Female s**	**Female s 95% C.I.**	**p-value (s)**	**Male μ**	**Male μ 95% C.I.**	**Female μ**	**Female μ 95% C.I.**	**p-value (μ)**
1q-	10	9	16.53	14.08 - 48.37	-	-	-	0.39	0.09 - 0.91	-	-	-
2p-	42	64	14.53	13.47 - 22.65	14.30	13.47 - 16.59	3.4×10^{-1}	2.12	0.50 - 3.68	3.41	1.34 - 5.16	1.8×10^{-1}
2q-	19	15	20.47	16.53 - 48.37	20.14	15.71 - 48.37	4.9×10^{-1}	0.28	0.15 - 0.48	0.24	0.12 - 0.51	3.8×10^{-1}
3p-	11	15	22.23	16.33 - 48.47	28.15	17.09 - 48.47	4.7×10^{-1}	0.16	0.09 - 0.35	0.15	0.10 - 0.27	5.2×10^{-1}
3q-	5	10	-	-	14.14	13.27 - 48.37	-	-	-	0.46	0.06 - 0.81	-
4q-	40	45	15.46	14.08 - 46.73	17.35	15.71 - 48.37	2.8×10^{-1}	1.49	0.37 - 2.43	0.90	0.37 - 1.3	2.1×10^{-1}
5q-	35	86	15.94	14.9 - 47.55	13.53	12.98 - 14.45	3.1×10^{-3}	0.56	0.21 - 0.80	2.26	1.46 - 2.96	1.9×10^{-5}
6p-	7	11	-	-	19.37	15.71 - 48.37	-	-	-	0.14	0.06 - 0.26	-
6q-	20	13	13.63	12.29 - 46.57	13.12	12.29 - 48.29	6.1×10^{-1}	0.54	0.10 - 0.87	0.41	0.05 - 0.75	4.0×10^{-1}
7p-	6	18	-	-	18.13	15.71 - 48.37	-	-	-	0.21	0.09 - 0.3	-
7q-	33	32	15.25	14.08 - 45.92	13.25	12.45 - 16.12	4.9×10^{-2}	0.74	0.21 - 1.14	0.91	0.33 - 1.5	2.6×10^{-1}
8p-	12	8	15.68	14.08 - 48.37	-	-	-	0.22	0.06 - 0.33	-	-	-
9q-	11	17	17.36	14.9 - 48.37	14.41	13.27 - 48.37	3.3×10^{-1}	0.14	0.05 - 0.22	0.60	0.10 - 1.05	7.2×10^{-2}
10q-	55	197	11.81	11.14 - 12.86	11.39	11.07 - 11.75	1.4×10^{-1}	2.12	1.21 - 2.98	8.15	6.42 - 9.99	3.5×10^{-8}
11p-	16	12	13.37	12.45 - 47.55	16.15	14.90 - 48.37	2.6×10^{-1}	0.92	0.11 - 1.77	0.26	0.07 - 0.42	1.2×10^{-1}
11q-	118	60	13.38	12.92 - 14.08	12.06	11.47 - 12.94	9.0×10^{-3}	3.23	2.35 - 4.13	2.14	1.29 - 3.05	6.9×10^{-2}
12q-	11	13	15.71	14.08 - 48.37	13.35	12.45 - 48.37	3.6×10^{-1}	0.22	0.06 - 0.34	0.81	0.07 - 1.60	1.5×10^{-1}
13q-	195	142	15.58	14.86 - 16.94	16.35	15.35 - 19.16	1.6×10^{-1}	5.25	3.52 - 6.47	2.32	1.38 - 2.97	5.1×10^{-4}
14q-	38	30	14.93	14.08 - 44.29	18.36	16.53 - 48.37	1.3×10^{-1}	1.16	0.29 - 1.76	0.4	0.19 - 0.58	4.6×10^{-2}
16p-	29	75	15.36	14.08 - 48.37	14.87	14.14 - 17.86	1.2×10^{-1}	1.77	0.35 - 3.02	3.76	1.36 - 5.85	1.1×10^{-1}
16q-	17	11	27.96	17.35 - 48.37	14.47	13.27 - 48.37	3.5×10^{-1}	0.21	0.15 - 0.39	0.29	0.05 - 0.49	6.1×10^{-1}
17q-	18	26	14.01	13.27 - 47.55	14.53	13.27 - 47.55	4.8×10^{-1}	1.70	0.17 - 3.01	2.19	0.25 - 4.66	4.6×10^{-1}
20q-	241	123	14.39	13.98 - 15.02	13.84	13.36 - 14.64	1.2×10^{-1}	7.77	5.93 - 9.40	4.22	2.84 - 5.55	2.0×10^{-3}
21q-	9	13	-	-	14.30	13.27 - 48.37	-	-	- - -	0.19	0.04 - 0.28	-
22q-	27	33	18.67	16.53 - 48.37	15.86	14.90 - 47.55	3.2×10^{-1}	0.47	0.23 - 0.70	0.79	0.23 - 1.22	2.9×10^{-1}

Table B.6 Sex-specific fitness effects and mutation rates for CNLOH events. The fitness effects and mutation rates were only calculated if the mCA was observed at least 10 times. Fitness effects and mutation rates were only calculated using data from individuals who had a single mCA. The 'observed number' refers to the number of individuals who had the mCA as their only mCA. p-values were calculated from the area under the distribution of difference probability curve where the difference ≤ 0.

	Observed number		**Fitness effect (s) (% per year)**					**mCA-specific mutation rate (μ) ($\times 10^{-9}$/ year)**				
mCA	**Men**	**Women**	**Male s**	**Male s 95% C.I.**	**Female s**	**Female s 95% C.I.**	**p-value (s)**	**Male μ**	**Male μ 95% C.I.**	**Female μ**	**Female μ 95% C.I.**	**p-value (μ)**
1=	25	39	9.99	9.02 - 12.33	12.21	11.43 - 14.29	3.4×10^{-2}	0.95	0.33 - 1.59	0.57	0.30 - 0.82	1.2×10^{-1}
1p=	274	314	12.02	11.69 - 12.38	12.63	12.30 - 12.99	1.1×10^{-2}	6.21	5.14 - 7.19	5.06	4.30 - 5.69	4.1×10^{-2}
1q=	208	224	11.41	11.07 - 11.81	11.37	11.03 - 11.72	4.4×10^{-1}	5.61	4.59 - 6.58	4.92	4.01 - 5.77	1.7×10^{-1}
2p=	42	53	12.35	11.63 - 14.08	10.98	10.33 - 11.92	2.3×10^{-2}	0.99	0.51 - 1.46	1.55	0.89 - 2.17	9.2×10^{-2}
2q=	61	78	11.53	10.90 - 12.53	10.95	10.39 - 11.69	1.3×10^{-1}	1.72	1.05 - 2.40	2.30	1.53 - 3.16	1.5×10^{-1}
3p=	55	51	11.11	10.44 - 12.07	10.14	9.47 - 11.12	6.0×10^{-2}	1.72	1.03 - 2.42	2.02	1.15 - 3.11	3.0×10^{-1}
3q=	48	44	10.62	9.92 - 11.76	11.01	10.35 - 12.26	2.8×10^{-1}	2.07	1.13 - 3.09	1.32	0.72 - 1.92	1.1×10^{-1}
4=	12	7	10.13	8.76 - 46.49	-	-	-	0.57	0.05 - 1.35	-	-	-
4p=	18	21	12.76	11.51 - 47.49	11.89	10.57 - 39.71	2.3×10^{-1}	0.53	0.10 - 0.91	0.52	0.09 - 0.91	5.7×10^{-1}
4q=	73	88	16.83	15.73 - 47.77	13.96	13.29 - 15.53	7.1×10^{-3}	0.76	0.43 - 0.96	1.07	0.70 - 1.35	6.1×10^{-2}
5q=	55	54	12.93	12.14 - 14.57	11.26	10.60 - 12.35	9.3×10^{-3}	0.94	0.55 - 1.29	1.50	0.88 - 2.15	5.9×10^{-2}
6p=	94	117	11.45	10.96 - 12.13	11.73	11.27 - 12.34	2.5×10^{-1}	2.92	2.01 - 3.76	2.79	2.08 - 3.53	4.1×10^{-1}
6q=	26	29	12.24	11.40 - 17.09	9.97	9.21 - 11.64	1.3×10^{-2}	0.68	0.21 - 1.05	1.90	0.66 - 3.54	2.6×10^{-2}
7p=	24	35	13.82	12.45 - 46.73	11.97	11.20 - 13.96	3.4×10^{-2}	0.37	0.11 - 0.55	0.80	0.38 - 1.21	3.0×10^{-2}
7q=	43	52	11.89	11.14 - 13.43	11.51	10.82 - 12.61	2.8×10^{-1}	1.04	0.57 - 1.46	1.52	0.85 - 2.17	1.3×10^{-1}
8p=	13	18	13.42	12.45 - 48.37	14.46	13.27 - 48.37	4.7×10^{-1}	0.25	0.06 - 0.38	0.22	0.07 - 0.31	4.9×10^{-1}
8q=	44	40	12.23	11.43 - 13.86	11.48	10.71 - 13.00	1.7×10^{-1}	1.43	0.71 - 2.11	1.11	0.57 - 1.70	2.8×10^{-1}
9=	26	33	10.15	9.29 - 12.41	8.39	7.71 - 9.71	2.2×10^{-2}	0.88	0.33 - 1.50	2.26	0.88 - 3.94	3.3×10^{-2}
9p=	150	125	15.77	14.84 - 18.69	13.74	13.07 - 14.87	5.9×10^{-3}	2.34	1.53 - 2.96	2.18	1.53 - 2.77	4.2×10^{-1}
9q=	128	158	12.13	11.64 - 12.71	13.25	12.73 - 13.96	4.9×10^{-3}	3.06	2.32 - 3.80	2.43	1.86 - 2.97	9.7×10^{-2}
10p=	18	19	11.90	10.51 - 44.73	12.15	11.39 - 44.73	4.6×10^{-1}	0.45	0.08 - 0.77	0.36	0.07 - 0.58	3.6×10^{-1}
10q=	28	46	11.69	10.80 - 14.39	12.22	11.43 - 13.68	3.4×10^{-1}	0.80	0.31 - 1.30	1.09	0.57 - 1.62	2.5×10^{-1}
11=	12	3	10.57	9.63 - 47.37	-	-	-	0.37	0.04 - 0.67	-	-	-
11p=	223	229	13.58	13.14 - 14.21	13.24	12.84 - 13.8	1.9×10^{-1}	3.90	3.09 - 4.55	3.76	3.02 - 4.44	4.0×10^{-1}
11q=	187	159	12.32	11.91 - 12.83	12.64	12.21 - 13.23	1.8×10^{-1}	4.31	3.42 - 5.09	2.83	2.18 - 3.45	4.6×10^{-3}
12p=	10	25	8.94	7.76 - 47.24	12.38	11.63 - 36.94	4.5×10^{-1}	1.32	0.04 - 4.22	0.58	0.12 - 0.91	1.1×10^{-1}
12q=	91	95	12.93	12.32 - 13.95	11.65	11.13 - 12.36	5.5×10^{-3}	1.58	1.09 - 1.99	2.15	1.52 - 2.75	7.4×10^{-2}
13q=	196	184	12.58	12.16 - 13.08	13.44	12.92 - 14.16	1.3×10^{-2}	4.26	3.42 - 5.02	2.59	2.03 - 3.02	7.1×10^{-4}
14q=	299	337	12.13	11.83 - 12.48	12.58	12.28 - 12.93	4.0×10^{-2}	6.41	5.47 - 7.30	5.39	4.57 - 6.04	6.2×10^{-2}
15q=	189	194	11.82	11.44 - 12.24	10.15	9.82 - 10.53	7.5×10^{-9}	4.14	3.33 - 4.81	8.21	6.34 - 9.93	1.1×10^{-5}
16=	16	24	9.76	8.42 - 21.37	10.33	9.47 - 12.65	4.4×10^{-1}	0.81	0.09 - 1.67	0.86	0.28 - 1.48	5.1×10^{-1}
16p=	105	117	11.01	10.55 - 11.59	12.7	12.22 - 13.45	5.6×10^{-5}	4.38	3.12 - 5.58	2.32	1.69 - 2.86	2.0×10^{-3}
16q=	84	87	11.31	10.80 - 12.02	11.39	10.86 - 12.14	4.3×10^{-1}	2.67	1.81 - 3.48	2.10	1.44 - 2.77	1.6×10^{-1}
17p=	42	42	12.93	12.14 - 15.57	12.96	12.14 - 15.86	4.8×10^{-1}	0.91	0.40 - 1.30	0.95	0.40 - 1.40	4.6×10^{-1}
17q=	139	166	11.84	11.42 - 12.40	12.66	12.21 - 13.29	1.5×10^{-2}	3.77	2.86 - 4.65	2.84	2.19 - 3.40	5.2×10^{-2}
18p=	10	4	10.51	9.63 - 47.37	-	-	-	0.78	0.04 - 1.66	-	-	-
18q=	25	45	11.40	10.43 - 14.29	12.14	11.35 - 13.67	2.9×10^{-1}	0.85	0.29 - 1.50	1.17	0.60 - 1.72	2.6×10^{-1}
19p=	56	83	11.33	10.73 - 12.30	10.98	10.49 - 11.65	2.3×10^{-1}	2.29	1.35 - 3.30	2.90	1.89 - 3.86	2.2×10^{-1}
19q=	81	78	12.73	12.10 - 13.69	11.78	11.22 - 12.61	4.0×10^{-2}	1.96	1.30 - 2.55	2.30	1.51 - 3.07	2.6×10^{-1}
20p=	15	23	11.93	10.67 - 46.65	11.5	10.63 - 15.86	2.2×10^{-1}	0.54	0.07 - 1.04	0.88	0.19 - 1.69	2.4×10^{-1}
20q=	68	75	12.71	12.02 - 13.96	11.99	11.40 - 12.87	1.0×10^{-1}	1.43	0.92 - 1.89	1.79	1.19 - 2.36	2.0×10^{-1}
21q=	62	69	10.71	10.12 - 11.55	11.97	11.40 - 12.87	1.1×10^{-2}	3.37	2.02 - 4.86	1.82	1.16 - 2.51	2.3×10^{-2}
22q=	129	163	13.65	13.01 - 14.66	14.63	13.92 - 15.96	6.7×10^{-2}	2.38	1.69 - 2.96	2.12	1.53 - 2.59	2.7×10^{-1}

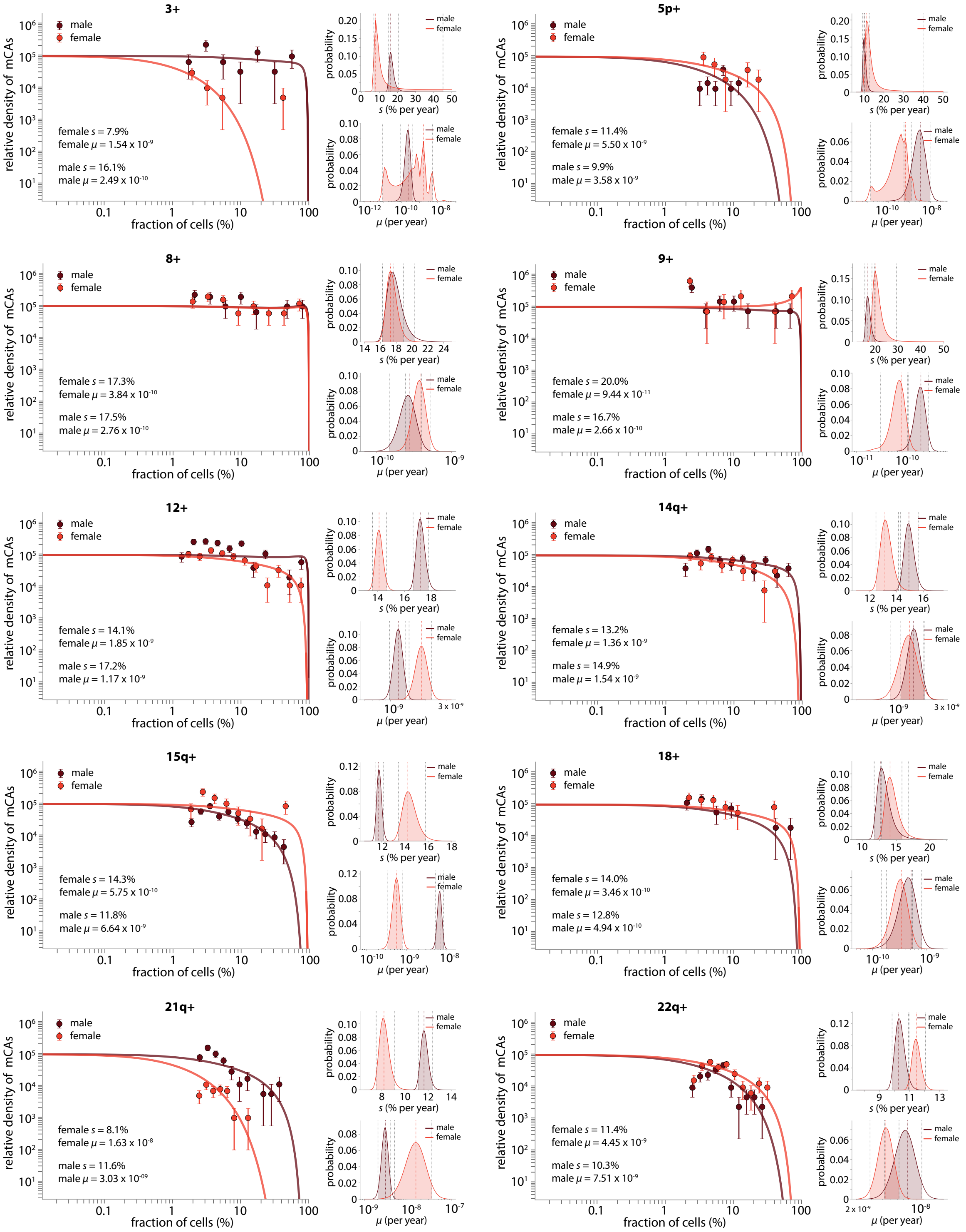


Figure B.12 Sex differences in fitness effects and mutation rates: gains. Only gain events which were observed 10 or more times in men (with a single mCA) and 10 or more times in women (with a single mCA) are shown. Shaded area, between the grey dashed vertical lines on the small subplots indicates the 95% confidence interval for the estimated s and μ values. The coloured vertical dashed line indicates the most likely s and μ values.

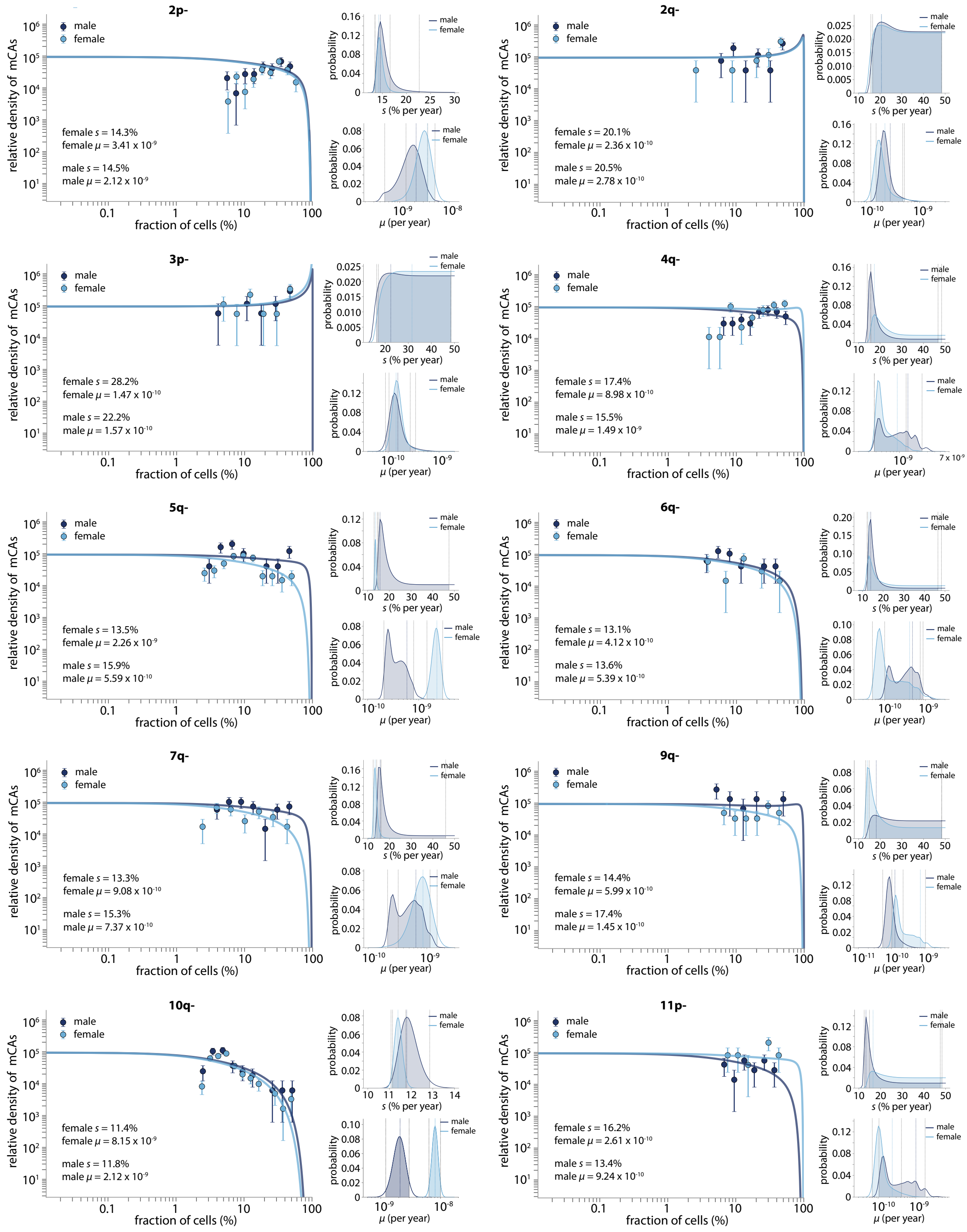


Figure B.13 Sex differences in fitness effects and mutation rates: losses: part 1. Only loss events which were observed 10 or more times in men (with a single mCA) and 10 or more times in women (with a single mCA) are shown. Shaded area, between the grey dashed vertical lines on the small subplots indicates the 95% confidence interval for the estimated s and μ values. The coloured vertical dashed line indicates the most likely s and μ values.

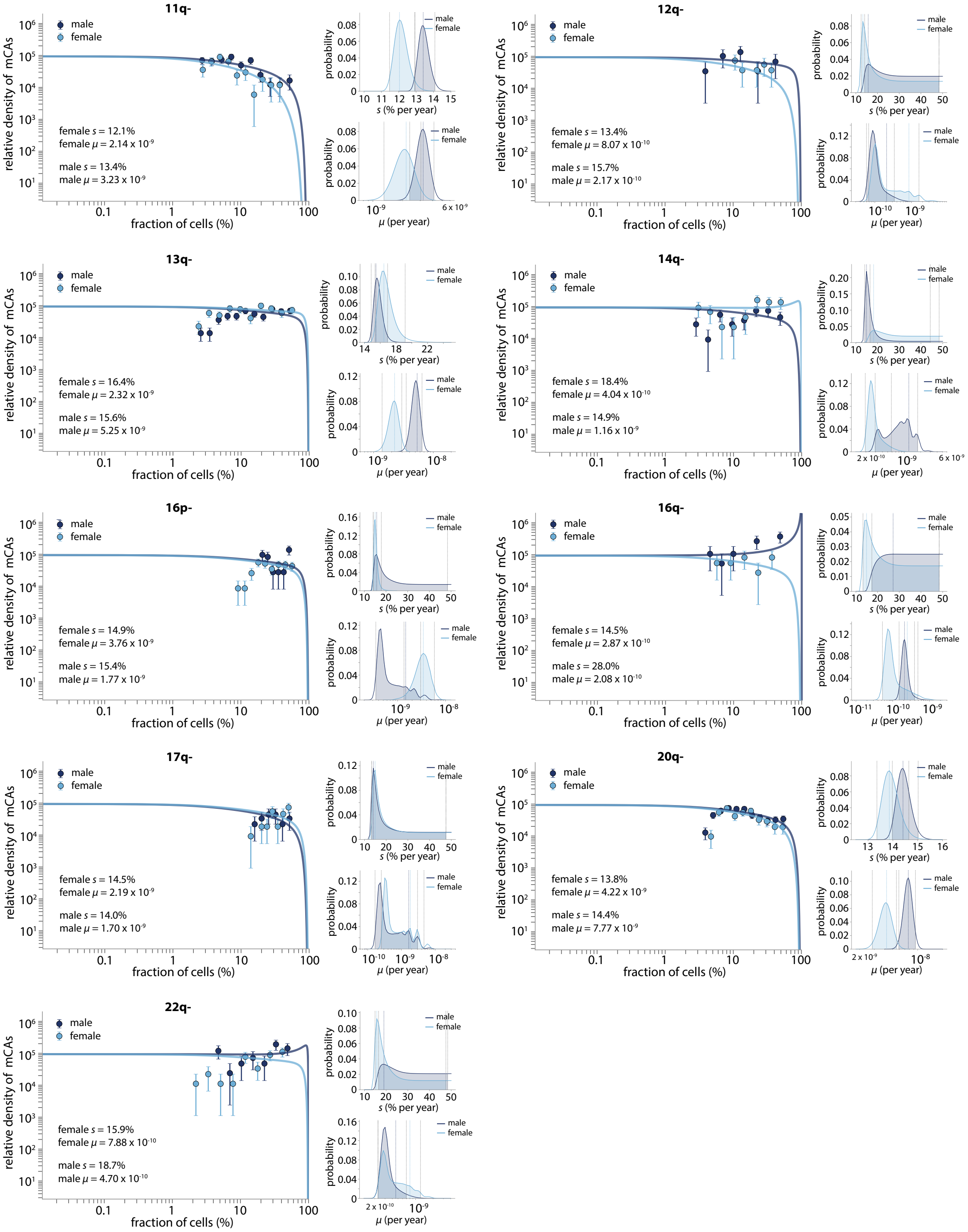


Figure B.14 Sex differences in fitness effects and mutation rates: losses: part 2. Only loss events which were observed 10 or more times in men (with a single mCA) and 10 or more times in women (with a single mCA) are shown. Shaded area, between the grey dashed vertical lines on the small subplots indicates the 95% confidence interval for the estimated s and μ values. The coloured vertical dashed line indicates the most likely s and μ values.

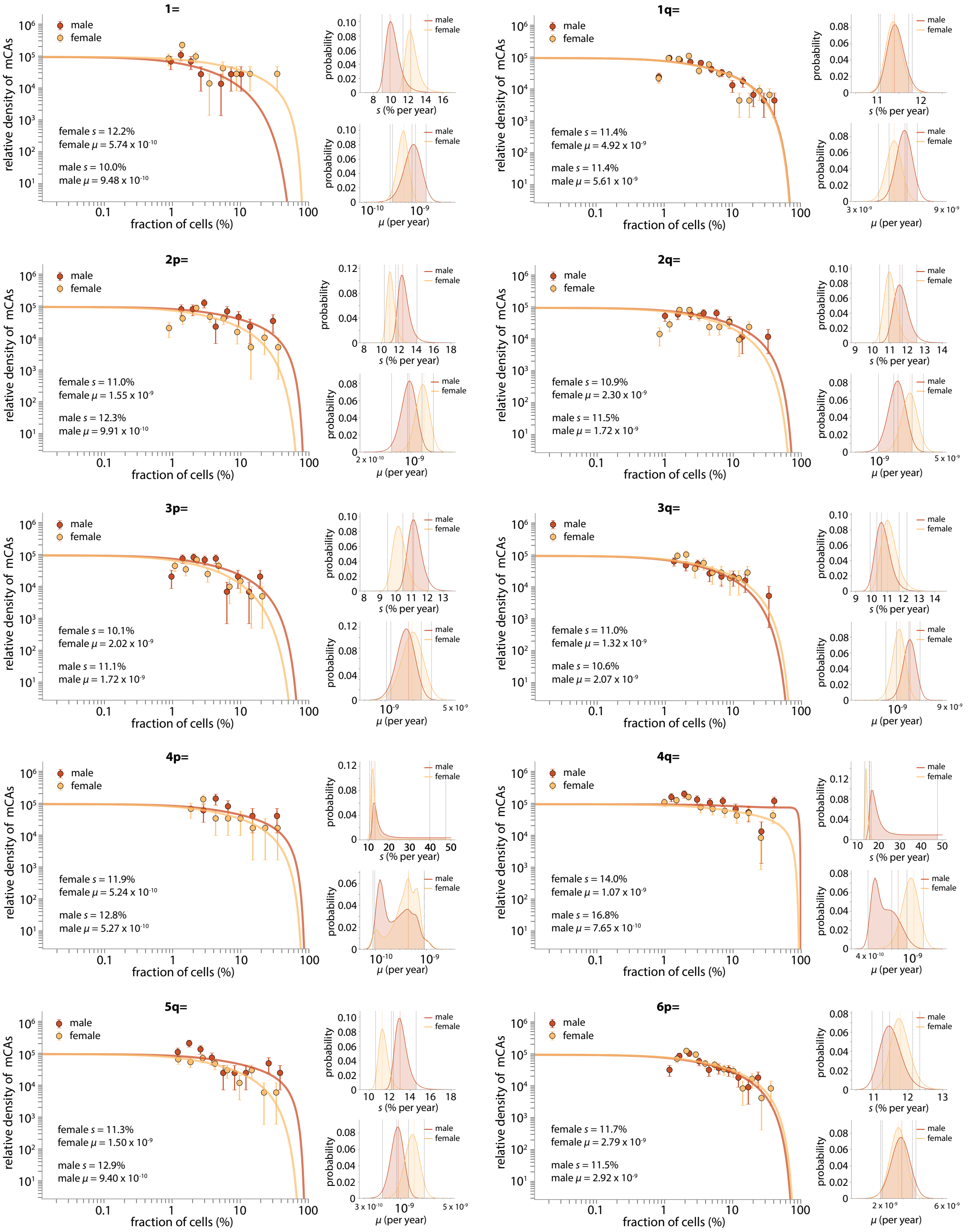


Figure B.15 Sex differences in fitness effects and mutation rates: CNLOH: part 1. Only CNLOH events which were observed 10 or more times in men (with a single mCA) and 10 or more times in women (with a single mCA) are shown. Shaded area, between the grey dashed vertical lines on the small subplots indicates the 95% confidence interval for the estimated s and μ values. The coloured vertical dashed line indicates the most likely s and μ values.

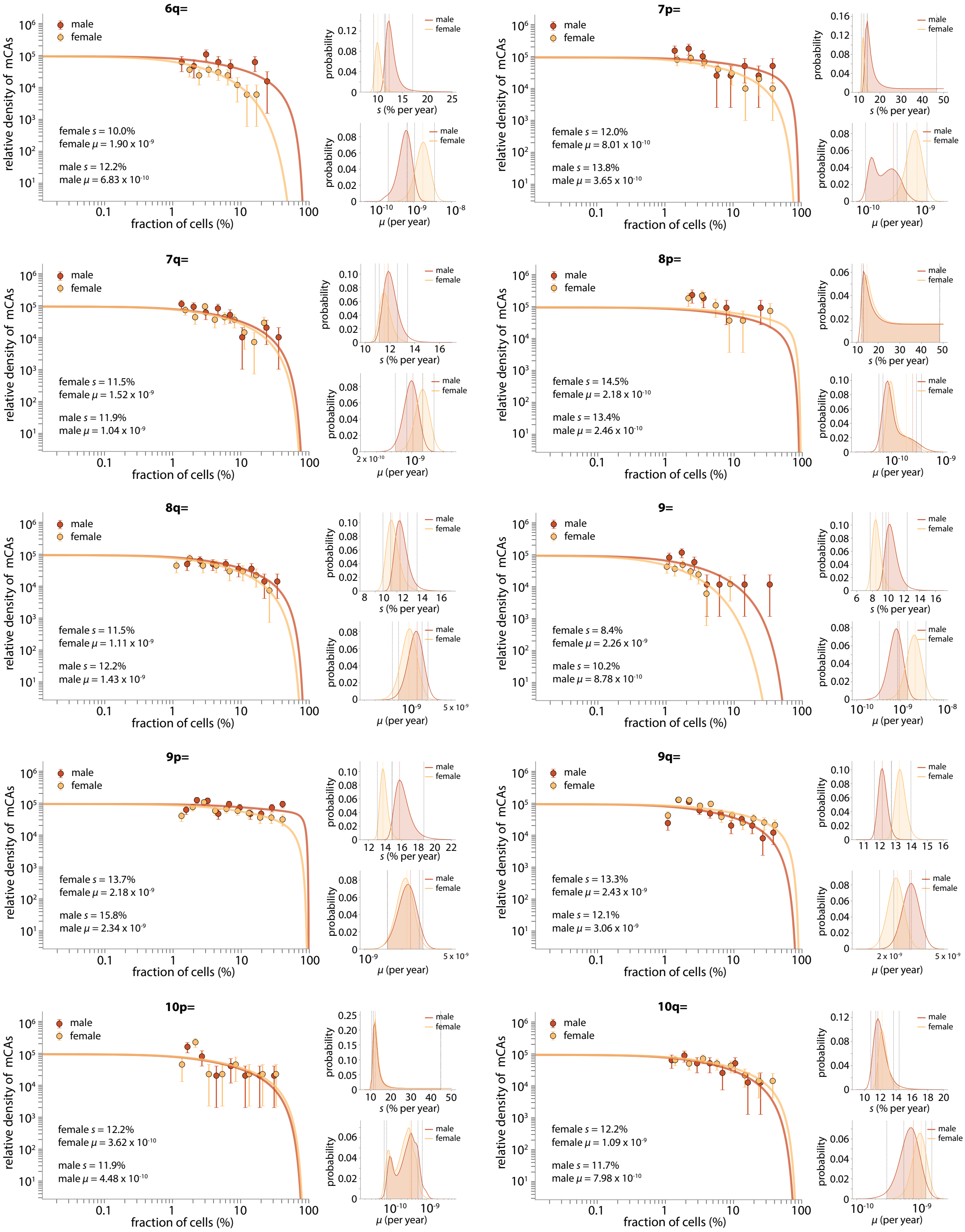


Figure B.16 Sex differences in fitness effects and mutation rates: CNLOH: part 2. Only CNLOH events which were observed 10 or more times in men (with a single mCA) and 10 or more times in women (with a single mCA) are shown. Shaded area, between the grey dashed vertical lines on the small subplots indicates the 95% confidence interval for the estimated s and μ values. The coloured vertical dashed line indicates the most likely s and μ values.

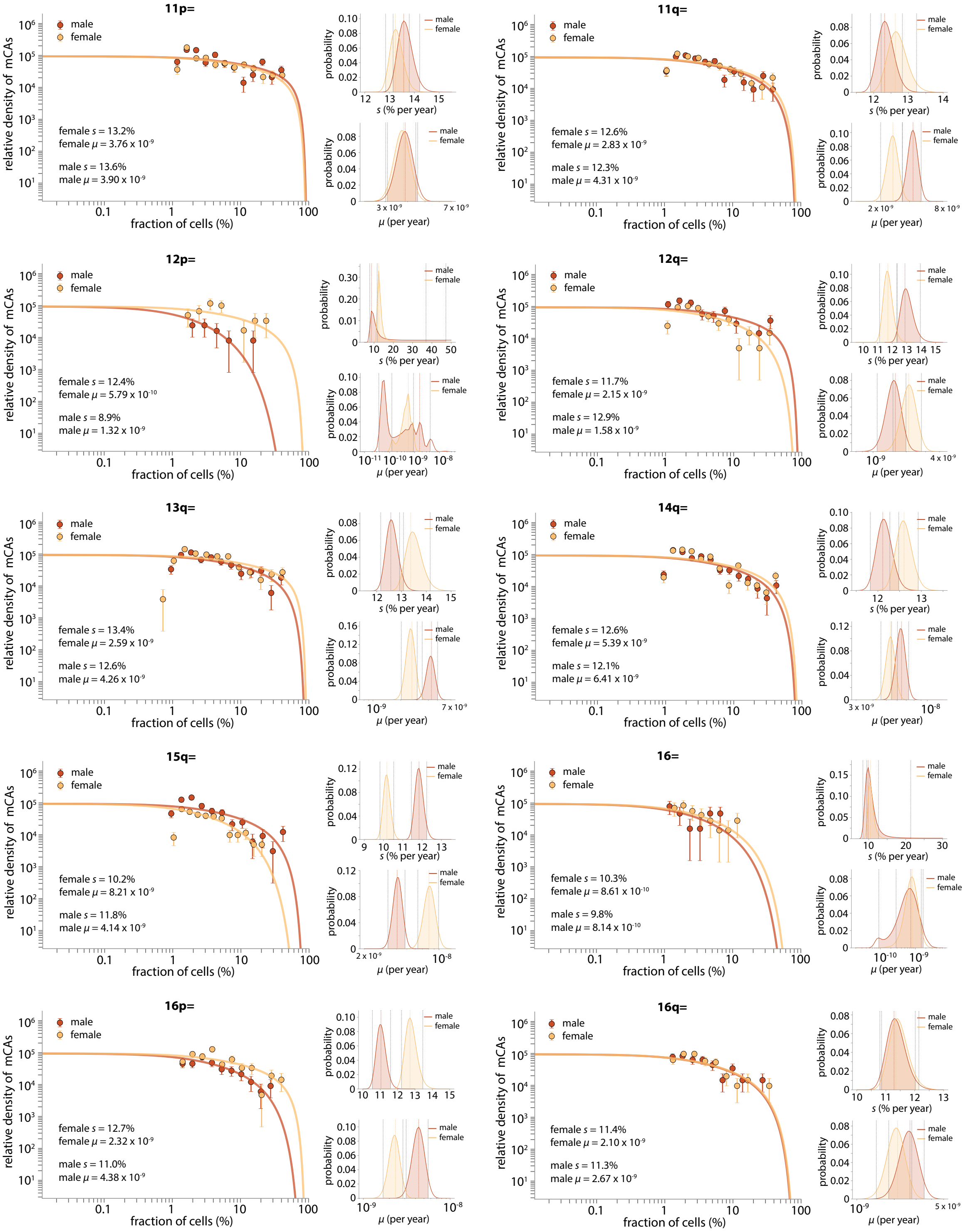


Figure B.17 Sex differences in fitness effects and mutation rates: CNLOH: part 3. Only CNLOH events which were observed 10 or more times in men (with a single mCA) and 10 or more times in women (with a single mCA) are shown. Shaded area, between the grey dashed vertical lines on the small subplots indicates the 95% confidence interval for the estimated s and μ values. The coloured vertical dashed line indicates the most likely s and μ values.

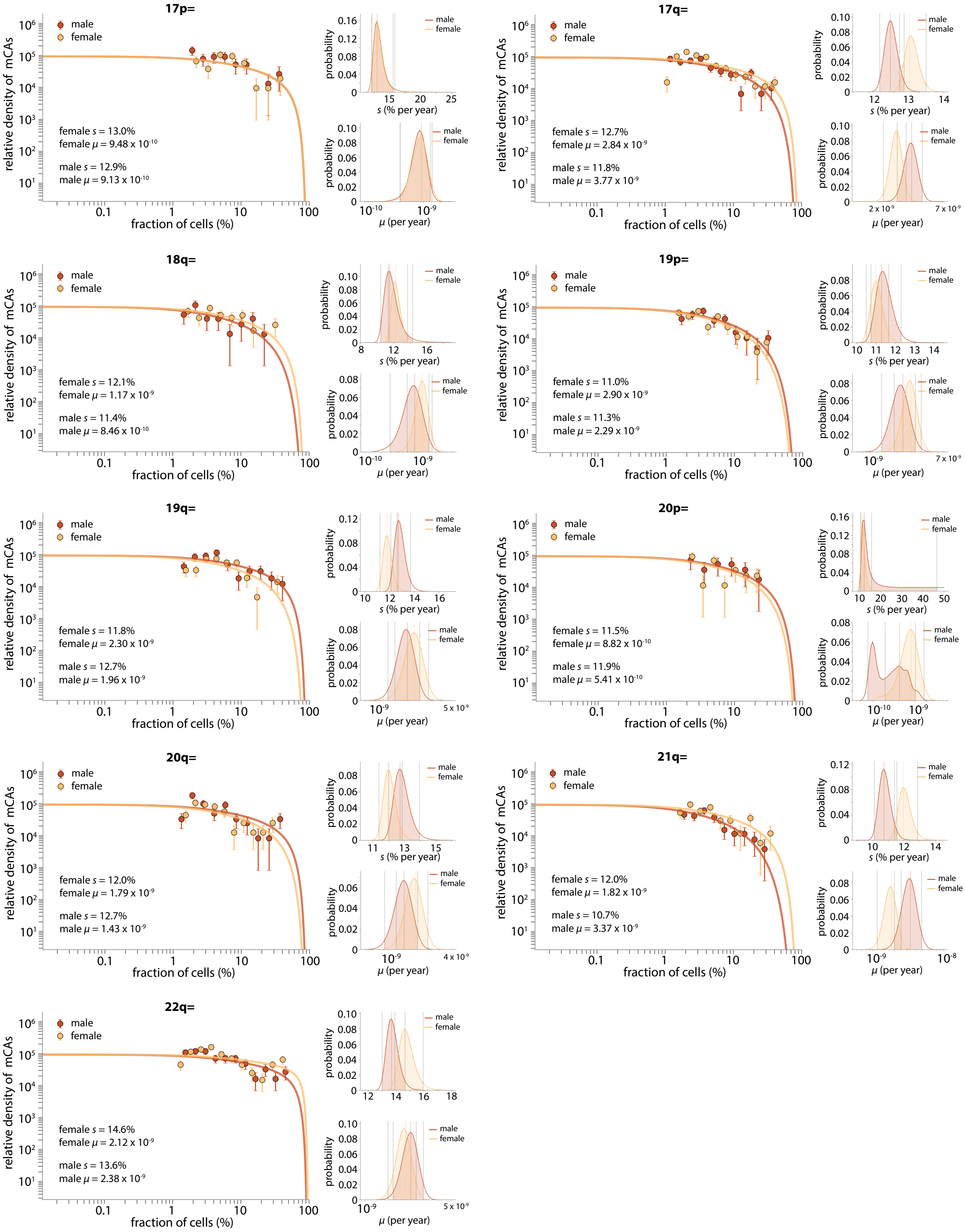


Figure B.18 **Sex differences in fitness effects and mutation rates: CNLOH: part 4**. Only CNLOH events which were observed 10 or more times in men (with a single mCA) and 10 or more times in women (with a single mCA) are shown. Shaded area, between the grey dashed vertical lines on the small subplots indicates the 95% confidence interval for the estimated s and μ values. The coloured vertical dashed line indicates the most likely s and μ values.

B.4 Age dependence of individual mCAs

Supplementary material for Section 3.5

After a threshold age determined by the limit of detection, the prevalence of a specific mCA is expected to increase roughly linearly at a rate determined by the mCA's fitness effect (reference Supplementary material 5: age dependence of individual mCAs). The limits of detection are different for each class of mCA, and even within a class, the limits of detection appear to be different (likely due to length of mCA, amongst other factors). The 'limit of detection' used for the age dependence plots was the lowest cell fraction observed for that mCA, multiplied by 1.5 (to try to take in to account the higher false negative rate at the lowest cell fractions).

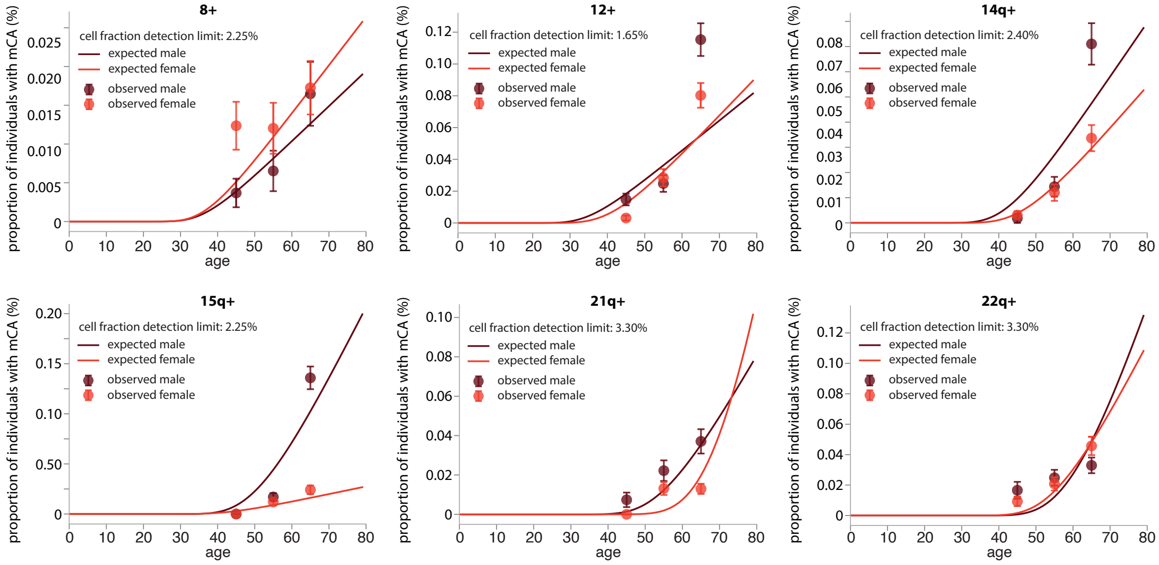


Figure B.19 Predicted age dependence for gain events calculated using sex-specific μ and s estimtes. Only gain events which were observed 30 or more times in both men and women are shown. The cell fraction limit of detection used was the minimum cell fraction observed for the mCA, multiplied by 1.5.

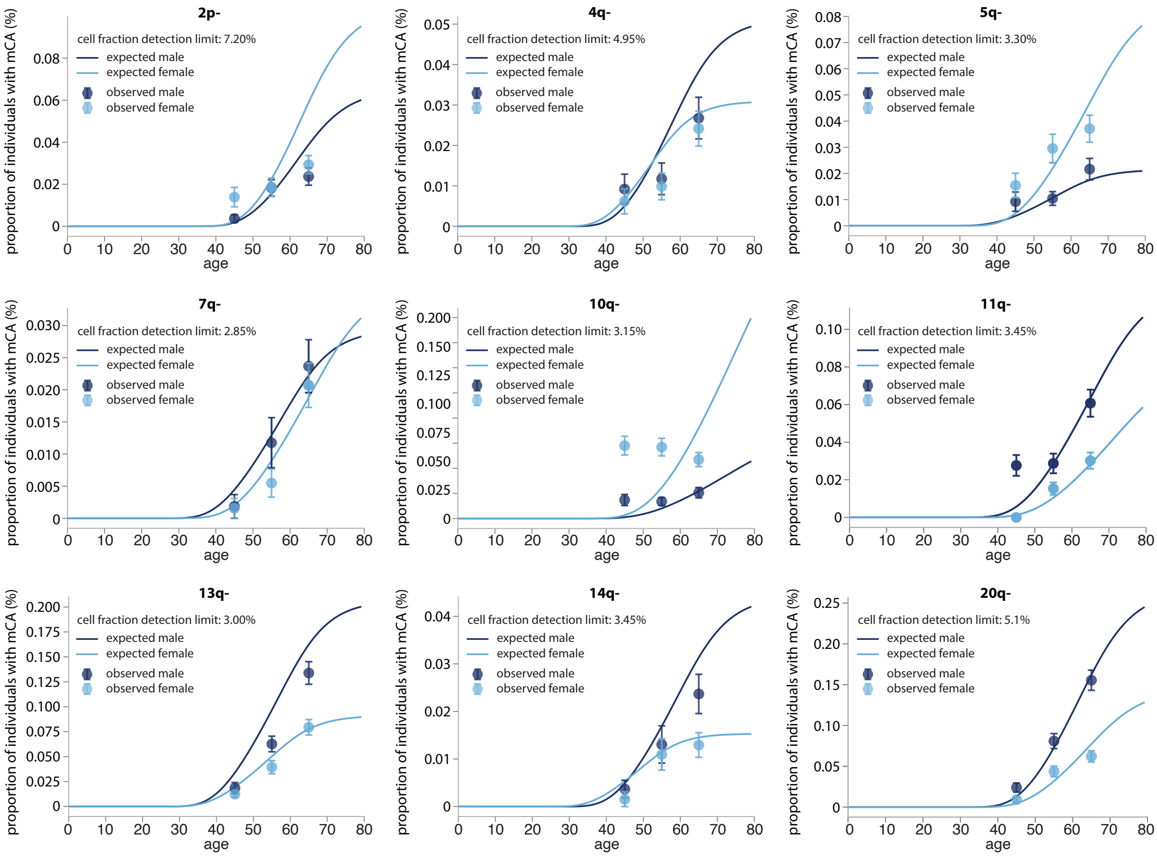


Figure B.20 Predicted age dependence for loss events calculated using sex-specific μ and s estimates. Only loss events which were observed 30 or more times in both men and women are shown. The cell fraction limit of detection used was the minimum cell fraction observed for the mCA, multiplied by 1.5.

.

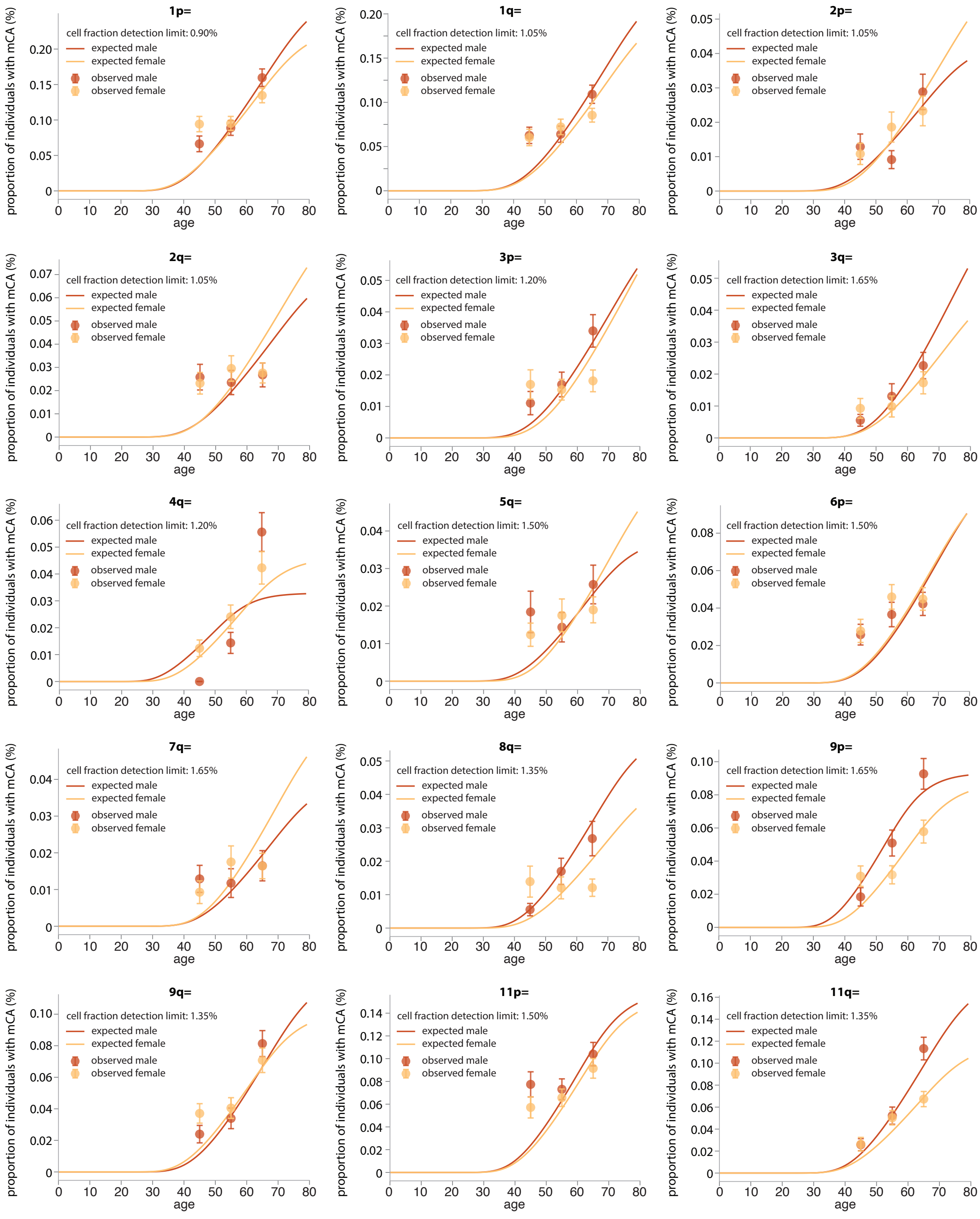


Figure B.21 Predicted age dependence for CNLOH events calculated using sex-specific μ and s estimates: part 1. Only CNLOH events which were observed 30 or more times in both men and women are shown. The cell fraction limit of detection used was the minimum cell fraction observed for the mCA, multiplied by 1.5.

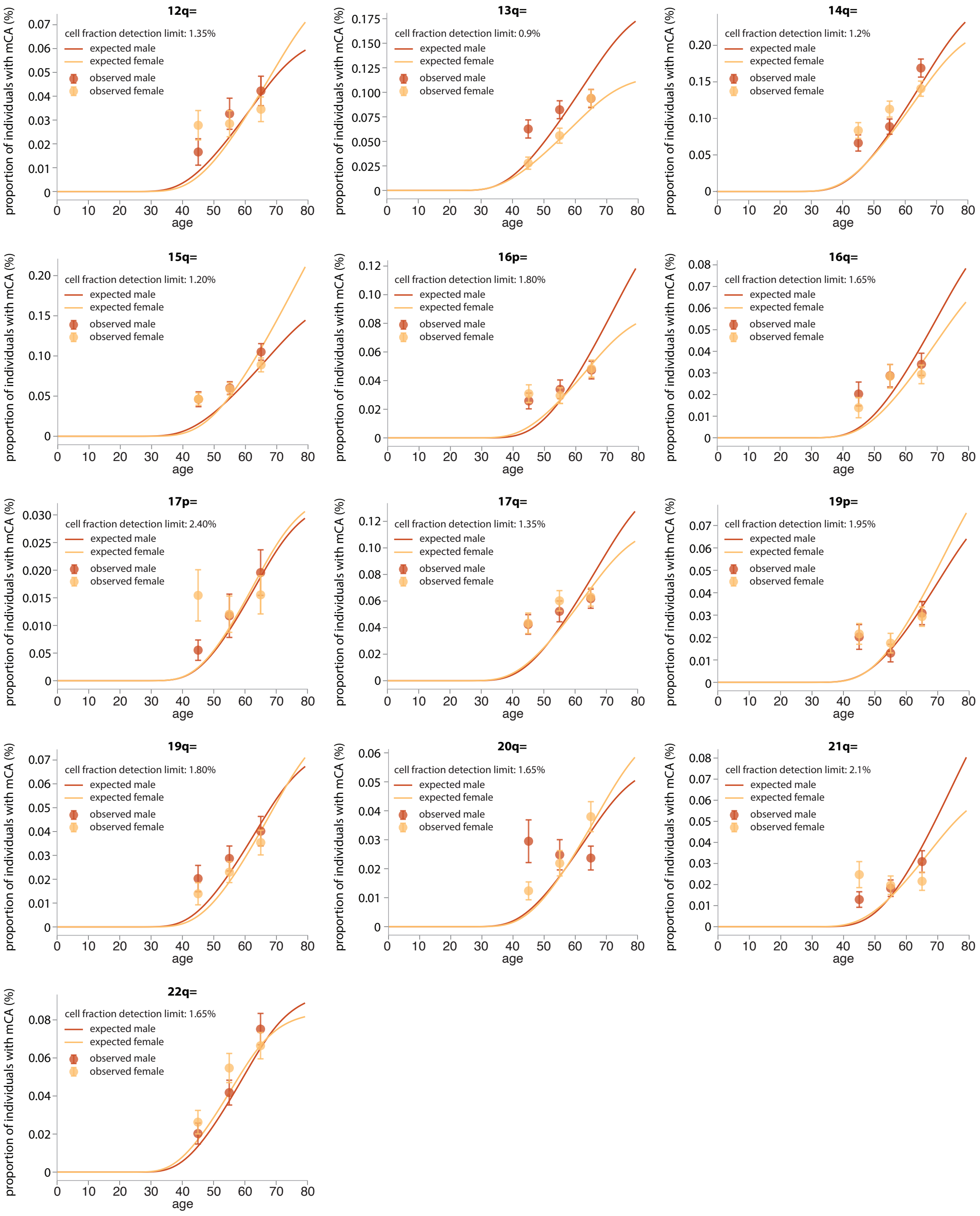


Figure B.22 Predicted age dependence for CNLOH events calculated using sex-specific μ and s estimates: part 2. Only CNLOH events which were observed 30 or more times in both men and women are shown. The cell fraction limit of detection used was the minimum cell fraction observed for the mCA, multiplied by 1.5.

B.4.1 Exploring mCAs that show deviation from expected age dependence

Supplementary material for Section 3.9

Can decline in prevalence with age be explained by acquisition of additional mCAs?

Several mCAs (10q-, 2q=, 3p= (women), 7q= (women), 8q= (women), 17p= (women), 20q= (men), 21q= (women)), seem to have a flat, or even decreasing, prevalence with increasing age. Could this be because individuals with these mCAs are more likely to acquire additional mCAs with increasing age, resulting in a decline in prevalence of the 'single mCA' with age? To look at this, we looked at the prevalence of these mCAs in individuals that $\geq$ 1 mCA (if the cell fraction difference between the mCAs was >2 %) and compared this observed prevalence to the expected prevalence based on the mCAs inferred fitness effect and mutation rate (Figure B.23). The poor age dependence persists, suggesting the reason is not the acquisition of additional mCAs.

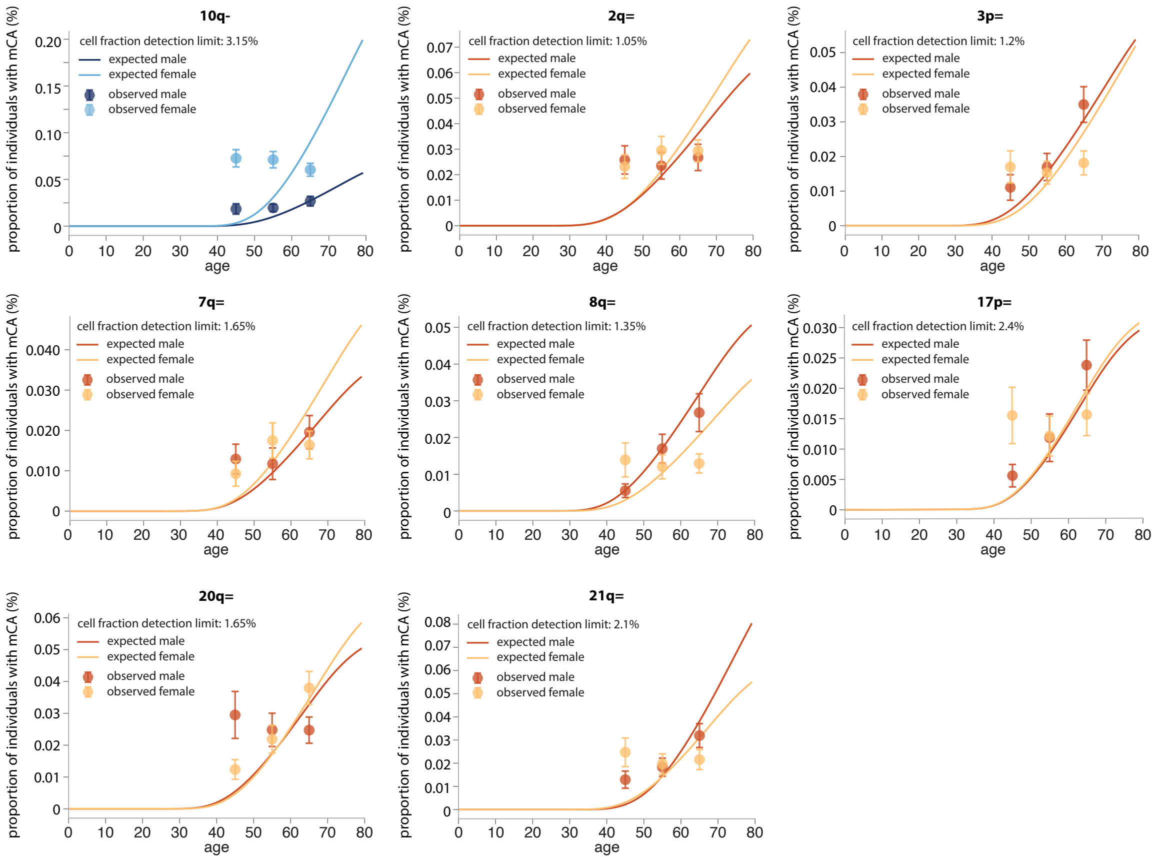


Figure B.23 Age and sex dependence of mCAs with poor age dependence, but including people with multiple mCAs. The cell fraction limit of detection used was the minimum cell fraction observed for the mCA, multiplied by 1.5. The predicted prevalence is for 'at least 1' mCA .

C

Supplementary material for Chapter 4

Table C.1 Genes commonly mutated in AML and clonal haematopoiesis. Genes are ranked in order of mutation prevalence within each AML or clonal haematopoiesis study. The percentage of the total number of variants found in the study which were attributable to that gene is also shown. Genes highlighted in red were chosen for our custom targeted sequencing panel.

Supplementary material for Section 4.2.1.

AML		Clonal haematopoiesis															
COSMIC v94[166]	Papaemmanuil 2016[20]	Jaiswal 2014[3]		Genovese 2014[4]		McKerrel 2015[6]		Zink 2017 (WGS)[9]		Acuna-Hidalgo 2017[8]		Coombs 2017[75]		Desai 2018[76]		Young 2016 & 2019[7,39]	
Gene	Gene	Gene	% variants	Gene	% variants	Gene	% variants	Gene	% variants	Gene	% variants	Gene	% variants	Gene	% variants	Gene	% variants
FLT3	FLT3	DNMT3A	49.08	DNMT3A	58.10	DNMT3A	41.96	DNMT3A	26.88	DNMT3A	60.54	DNMT3A	35.68	DNMT3A	57.14	DNMT3A	38.90
NPM1	NPM1	TET2	8.83	ASXL1	10.70	JAK2	22.32	TET2	25.14	CBL	5.80	TET2	11.07	TET2	12.70	TET2	15.60
DNMT3A	DNMT3A	ASXL1	7.61	TET2	9.48	SF3B1	16.07	GOLGA8B	8.38	TET2	4.93	ASXL1	5.45	ASXL1	4.76	CUX1	4.61
NRAS	NRAS	TP53	4.05	JAK2	7.34	SRSF2	13.39	ASXL1	7.23	GNAS	4.84	PPM1D	2.46	JAK1	3.17	BCORL1	3.96
TET2	TET2	JAK2	3.80	PPM1D	4.59	KRAS	1.79	PPM1D	5.20	TP53	4.04	ATM	2.11	JAK2	3.17	TP53	3.74
IDH2	IDH2	SF3B1	3.31	SF3B1	3.98	IDH1	1.70	TP53	5.20	JAK2	3.13	TP53	2.11	KDM6A	3.17	KDM6A	3.74
CEBPA	CEBPA	GNB1	2.70	SRSF2	2.14	IDH2	0.89	CUX1	2.60	ASXL1	1.79	SRSF2	1.05	SF3B1	3.17	ASXL1	2.42
KIT	RUNX1	CBL	1.47	TP53	1.22			JAK2	2.31	GNB1	1.79	STAG2	1.05	ATRX	1.59	STAG2	2.20
WT1	PTPN11	SRSF2	1.35	CBL	0.92			SF3B1	2.02	NRAS	1.79	CREBBP	0.88	BCOR	1.59	BCORL1	2.20
RUNX1	IDH1	GNAS	0.98	IDH2	0.31			SRSF2	1.73	KRAS	1.35	MLL3	0.88	CBL	1.59	CBL	1.98
ASXL1	TP53	CREBBP	0.98	MYD88	0.31			JHRC	1.73	PTPN11	1.35	STAT3	0.88	CREBBP	1.59	EZH2	1.98
TP53	SRSF2	NRAS	0.74	U2AF1	0.31			KMT2D	1.73	IDH2	0.90	JAK2	0.88	FLT1	1.59	NOTCH1	1.76
PTPN11	KMT2A	BRCC3	0.74	ARM	0.31			ACSS2	1.16	SRSF2	0.90	U2AF1	0.88	HRAS	1.59	ATRX	1.54
IDH1	WT1	RAD21	0.74	STAT3	0.31			AKAP17A	1.16	MYD88	0.90	EZH2	0.88	KIT	1.59	GATA2	1.32
KRAS	KRAS	U2AF1	0.61					CBL	0.87	BRAF	0.90	RAD21	0.88	NOTCH1	1.59	KRAS	1.32
SRSF2	ASXL1	STST3	0.49					SYNE1	0.87	SF3B1	0.90	AR	0.88			IKZF1	1.32
GATA2	KIT	BCOR	0.49					BCOR	0.58	SMAD4	0.45	CBL	0.88			KIT	0.88
STAG2	STAG2	CASP8	0.49					MYD88	0.58	RHEB	0.45	SF3B1	0.70			MYD88	0.88
RAD21	RAD21	MLL2	0.49					IDH2	0.58	ADNP	0.45	TERT	0.70			FBXW7	0.88
JAK2	EZH2	SETD2	0.49					NRAS	0.58	KCNQ3	0.45	EGFR	0.53			ETV6	0.88
	PHF6	KRAS	0.37					UMODL1	0.58	PIK3CA	0.45	MLL2	0.53			RUNX1	0.88
	SF3B1	IDH2	0.37					GNB1	0.29	HECTD1	0.45	KDM5C	0.53			SMC3	0.88
	CBL	CUX2	0.37					GNAS	0.29	CUX2	0.45	FAT1	0.53			RAD21	0.66
	U2AF1	PTPN11	0.37					CREBBP	0.29	COL4A3BP	0.45	NRAS	0.53			ABL1	0.66
	BCOR	LUC7L2	0.37					STAT3	0.29	BRCC3	0.45	FGFR1	0.53			ZRSR2	0.66
	GATA2	PRDM1	0.37					SETD2	0.29	U2AF1	0.45	PTPRS	0.53			KMT2A	0.44
	NH1	TNFAIP3	0.37					KIT	0.29			FLT3	0.53			PTPN11	0.44
	MYC	ZRSF2	0.37					ATM	0.29			SH2B3	0.53			CEBPA	0.44
	EP300	MYD88	0.25					CEP112	0.29			ARID1A	0.53			CSKN2A	0.44
	ETV6	BRAF	0.25					MPL	0.29			ARID1B	0.53			GNAS	0.22
	KDM5A	APC	0.25					ROBO1	0.29			BCOR	0.53			PTEN	0.22
	MLL2	BCORL1	0.25									SETD2	0.53			FLT3	0.22
	ZRSR2	DDX3X	0.25									CALR	0.53			CSF3R	0.22
	JAK2	EZH2	0.25									MET	0.35			SF3B1	0.22
	CREBBBP	KDM6A	0.25									GRIN2A	0.35			SRSF2	0.22
	KDM6A	NOTCH1	0.25									IDH2	0.35			CBLB	0.22
	MLL3	NOTCH2	0.25									WT1	0.35			CALR	0.22
	BRAF	SF1	0.25									IRS1	0.35			U2AF1	0.22
	FBXW7	TNFRSF14	0.25									RTSC12	0.35			WT1	0.22
	...	...										...				JAK2	0.22

Table C.2 Genomic coordinates of regions targeted in the custom SNV/ indel panel

Gene	Transcript	Chromosome	Location	Start	End
ASXL1	ENST00000375687.4	20	exon 11	31021086	31021720
	ENST00000375687.4	20	exon 12	31022234	31025141
BCOR	ENST00000378444.4	X	exon 2	39937096	39937182
	ENST00000378444.4	X	exon 3	39935705	39935785
	ENST00000378444.4	X	exon 4	39931601	39934433
	ENST00000378444.4	X	exon 5	39930889	39930943
	ENST00000378444.4	X	exon 6	39930225	39930412
	ENST00000378444.4	X	exon 7	39923588	39923852
	ENST00000378444.4	X	exon 8	39922860	39923205
	ENST00000378444.4	X	exon 9	39921998	39922324
	ENST00000378444.4	X	exon 10	39921391	39921646
	ENST00000378444.4	X	exon 11	39916407	39916574
	ENST00000378444.4	X	exon 12	39914620	39914766
	ENST00000378444.4	X	exon 13	39913508	39913586
	ENST00000378444.4	X	exon 14	39913137	39913295
	ENST00000378444.4	X	exon 15	39910500	39911653
BCORL1	ENST00000540052.1	X	exon 1	129139207	129139293
	ENST00000540052.1	X	exon 2	129146553	129146644
	ENST00000540052.1	X	exon 3	129146925	129150189
	ENST00000540052.1	X	exon 4	129154959	129155125
	ENST00000540052.1	X	exon 5	129156871	129156952
	ENST00000540052.1	X	exon 6	129158964	129159354
	ENST00000540052.1	X	exon 7	129162609	129162836
	ENST00000540052.1	X	exon 8	129171341	129171508
	ENST00000540052.1	X	exon 9	129173111	129173257
	ENST00000540052.1	X	exon 10	129184691	129184769
	ENST00000540052.1	X	exon 11	129185834	129185991
	ENST00000540052.1	X	exon 12	129189828	129190113
CBL	ENST00000264033.4	11	exon 8	119148875	119149007
	ENST00000264033.4	11	exon 9	119149219	119149423
	ENST00000264033.4	11	exon 16	119170204	119170491
CEBPA	ENST00000498907.2	19	exon 1	33792243	33793326
CHEK2	ENST00000328354.6	22	exon 1	29137756	29137832
	ENST00000328354.6	22	exon 2	29130390	29130715
	ENST00000328354.6	22	exon 3	29121230	29121355
	ENST00000328354.6	22	exon 4	29120964	29121112
	ENST00000328354.6	22	exon 5	29115382	29115473
	ENST00000328354.6	22	exon 6	29107896	29108005
	ENST00000328354.6	22	exon 7	29105993	29106047
	ENST00000328354.6	22	exon 8	29099492	29099554
	ENST00000328354.6	22	exon 9	29095825	29095925
	ENST00000328354.6	22	exon 10	29092888	29092975
	ENST00000328354.6	22	exon 11	29091697	29091861
	ENST00000328354.6	22	exon 12	29091114	29091230
	ENST00000328354.6	22	exon 13	29090019	29090105
	ENST00000328354.6	22	exon 14	29085122	29085203
	ENST00000328354.6	22	exon 15	29083731	29083974
CSF3R	ENST00000373106.1	1	exon 14	36933422	36933563
	ENST00000373106.1	1	exon 15	36933158	36933252
	ENST00000373106.1	1	exon 16	36932830	36932912
	ENST00000373106.1	1	exon 17	36931957	36932428
DDX41	ENST00000507955.1	5	exon 1	176943919	176943948
	ENST00000507955.1	5	exon 2	176943725	176943836
	ENST00000507955.1	5	exon 3	176943288	176943448
	ENST00000507955.1	5	exon 4	176943119	176943194
	ENST00000507955.1	5	exon 5	176942929	176942990
	ENST00000507955.1	5	exon 6	176942685	176942822
	ENST00000507955.1	5	exon 7	176942186	176942259
	ENST00000507955.1	5	exon 8	176941916	176942070
	ENST00000507955.1	5	exon 9	176941701	176941838
	ENST00000507955.1	5	exon 10	176940685	176940848
	ENST00000507955.1	5	exon 11	176940353	176940485
	ENST00000507955.1	5	exon 12	176940011	176940083
	ENST00000507955.1	5	exon 13	176939780	176939877
	ENST00000507955.1	5	exon 14	176939496	176939646
	ENST00000507955.1	5	exon 15	176939322	176939394
	ENST00000507955.1	5	exon 16	176939096	176939207
	ENST00000507955.1	5	exon 17	176938788	176938928
DNMT3A	ENST00000321117.5	2	exon 23	25457147	25457289
	ENST00000321117.5	2	exon 22	25458575	25458694
	ENST00000321117.5	2	exon 21	25459804	25459874
	ENST00000321117.5	2	exon 20	25461998	25462084

Gene	Transcript	Chromosome	Location	Start	End
DNMT3A (cont.)	ENST00000321117.5	2	exon 19	25463170	25463319
	ENST00000321117.5	2	exon 18	25463508	25463599
	ENST00000321117.5	2	exon 17	25464430	25464576
	ENST00000321117.5	2	exon 16	25466766	25466851
	ENST00000321117.5	2	exon 15	25467023	25467207
	ENST00000321117.5	2	exon 14	25467408	25467521
	ENST00000321117.5	2	exon 13	25468121	25468201
	ENST00000321117.5	2	exon 12	25468888	25468933
	ENST00000321117.5	2	exon 11	25469028	25469178
	ENST00000321117.5	2	exon 10	25469488	25469645
	ENST00000321117.5	2	exon 9	25469919	25470027
	ENST00000321117.5	2	exon 8	25470459	25470618
	ENST00000321117.5	2	exon 7	25470905	25471121
	ENST00000321117.5	2	exon 6	25497809	25497956
	ENST00000321117.5	2	exon 5	25498368	25498412
	ENST00000321117.5	2	exon 4	25505309	25505580
	ENST00000321117.5	2	exon 3	25523007	25523112
	ENST00000321117.5	2	exon 2	25536781	25536853
EZH2	ENST00000320356.2	7	exon 2	148544273	148544397
	ENST00000320356.2	7	exon 3	148543561	148543690
	ENST00000320356.2	7	exon 4	148529725	148529842
	ENST00000320356.2	7	exon 5	148526819	148526940
	ENST00000320356.2	7	exon 6	148525831	148525972
	ENST00000320356.2	7	exon 7	148524255	148524358
	ENST00000320356.2	7	exon 8	148523545	148523724
	ENST00000320356.2	7	exon 9	148516687	148516779
	ENST00000320356.2	7	exon 10	148514968	148515209
	ENST00000320356.2	7	exon 11	148514313	148514483
	ENST00000320356.2	7	exon 12	148513775	148513870
	ENST00000320356.2	7	exon 13	148512597	148512638
	ENST00000320356.2	7	exon 14	148512005	148512131
	ENST00000320356.2	7	exon 15	148511050	148511229
	ENST00000320356.2	7	exon 16	148508716	148508812
	ENST00000320356.2	7	exon 17	148507424	148507506
	ENST00000320356.2	7	exon 18	148506401	148506482
	ENST00000320356.2	7	exon 19	148506162	148506247
	ENST00000320356.2	7	exon 20	148504737	148504798
FLT3	ENST00000241453.7	13	exon 20	28592603	28592726
	ENST00000241453.7	13	exon 17	28601224	28601378
	ENST00000241453.7	13	exon 16	28602314	28602425
	ENST00000241453.7	13	exon 15	28608023	28608128
	ENST00000241453.7	13	exon 14	28608218	28608351
	ENST00000241453.7	13	exon 13	28608437	28608544
	ENST00000241453.7	13	exon 12	28609631	28609810
	ENST00000241453.7	13	exon 11	28610071	28610180
	ENST00000241453.7	13	exon 8	28623520	28623674
	ENST00000241453.7	13	exon 6	28624231	28624359
	ENST00000241453.7	13	exon 3	28636003	28636206
GATA2	ENST00000487848.1	3	exon 5	128202702	128202848
	ENST00000487848.1	3	exon 6	128200661	128200787
	ENST00000487848.1	3	exon 7	128199861	128200161
GNAS	ENST00000371100.4	20	exon 1	57427769	57430388
	ENST00000371100.4	20	exon 2	57470666	57470739
	ENST00000371100.4	20	exon 3	57473994	57474040
	ENST00000371100.4	20	exon 4	57478585	57478640
	ENST00000371100.4	20	exon 5	57478726	57478846
	ENST00000371100.4	20	exon 6	57480437	57480535
	ENST00000371100.4	20	exon 7	57484216	57484271
	ENST00000371100.4	20	exon 8	57484404	57484478
	ENST00000371100.4	20	exon 9	57484575	57484634
	ENST00000371100.4	20	exon 10	57484738	57484859
	ENST00000371100.4	20	exon 11	57485005	57485136
	ENST00000371100.4	20	exon 12	57485388	57485456
	ENST00000371100.4	20	exon 13	57485737	57486247
GNB1	ENST00000378609.4	1	exon 5	1747194	1747301
IDH1	ENST00000345146.2	2	exon 6	209108150	209108328
	ENST00000345146.2	2	exon 4	209113092	209113384
IDH2	ENST00000330062.3	15	exon 8	90628506	90628619
	ENST00000330062.3	15	exon 4	90631818	90631979
JAK2	ENST00000381652.3	9	exon 6	5050685	5050831
	ENST00000381652.3	9	exon 12	5069924	5070054
	ENST00000381652.3	9	exon 14	5073697	5073785
KIT	ENST00000288135.5	4	exon 1	55524181	55524248
	ENST00000288135.5	4	exon 2	55561677	55561947
	ENST00000288135.5	4	exon 7	55575589	55575705
	ENST00000288135.5	4	exon 8	55589749	55589864
	ENST00000288135.5	4	exon 9	55592022	55592216
	ENST00000288135.5	4	exon 10	55593383	55593490
	ENST00000288135.5	4	exon 11	55593581	55593708

Gene	Transcript	Chromosome	Location	Start	End
KIT (cont.)	ENST00000288135.5	4	exon 12	55593988	55594093
	ENST00000288135.5	4	exon 13	55594176	55594287
	ENST00000288135.5	4	exon 16	55598036	55598164
	ENST00000288135.5	4	exon 17	55599235	55599358
KRAS	ENST00000256078.4	12	exon 5	25368371	25368494
	ENST00000256078.4	12	exon 4	25378547	25378707
	ENST00000256078.4	12	exon 3	25380167	25380346
	ENST00000256078.4	12	exon 2	25398207	25398329
MPL	ENST00000372470.3	1	exon 9	43814513	43814673
	ENST00000372470.3	1	exon 10	43814933	43815030
	ENST00000372470.3	1	exon 11	43817886	43817974
	ENST00000372470.3	1	exon 12	43818188	43818443
NPM1	ENST00000296930.5	5	exon 11	170837530	170837569
NRAS	ENST00000369535.4	1	exon 3	115256420	115256599
	ENST00000369535.4	1	exon 2	115258670	115258781
PPM1D	ENST00000305921.3	17	exon 1	58677775	58678247
	ENST00000305921.3	17	exon 5	58733959	58734202
	ENST00000305921.3	17	exon 6	58740355	58740913
PTPN11	ENST00000351677.2	12	exon 3	112888121	112888316
	ENST00000351677.2	12	exon 7	112910747	112910844
	ENST00000351677.2	12	exon 8	112915454	112915534
	ENST00000351677.2	12	exon 13	112926827	112926979
RAD21	ENST00000297338.2	8	exon 2	117878824	117878969
	ENST00000297338.2	8	exon 3	117875368	117875498
	ENST00000297338.2	8	exon 4	117874079	117874179
	ENST00000297338.2	8	exon 5	117870590	117870697
	ENST00000297338.2	8	exon 6	117869505	117869712
	ENST00000297338.2	8	exon 7	117868884	117869010
	ENST00000297338.2	8	exon 8	117868404	117868527
	ENST00000297338.2	8	exon 9	117866483	117866707
	ENST00000297338.2	8	exon 10	117864787	117864947
	ENST00000297338.2	8	exon 11	117864186	117864335
	ENST00000297338.2	8	exon 12	117862856	117863006
	ENST00000297338.2	8	exon 13	117861184	117861268
	ENST00000297338.2	8	exon 14	117859737	117859930
RUNX1	ENST00000344691.4	21	exon 6	36164431	36164907
	ENST00000344691.4	21	exon 5	36171597	36171759
	ENST00000344691.4	21	exon 4	36206706	36206898
	ENST00000344691.4	21	exon 3	36231770	36231875
	ENST00000344691.4	21	exon 2	36252853	36253010
	ENST00000344691.4	21	exon 1	36259139	36259409
SF3B1	ENST00000335508.6	2	exon 24	198257695	198257912
	ENST00000335508.6	2	exon 18	198265438	198265660
	ENST00000335508.6	2	exon 16	198266465	198266612
	ENST00000335508.6	2	exon 15	198266708	198266854
	ENST00000335508.6	2	exon 14	198267279	198267550
	ENST00000335508.6	2	exon 13	198267672	198267759
	ENST00000335508.6	2	exon 6	198281464	198281635
	ENST00000335508.6	2	exon 5	198283232	198283312
	ENST00000335508.6	2	exon 4	198285151	198285266
	ENST00000335508.6	2	exon 3	198285752	198285857
SRSF2	ENST00000392485.2	17	exon 2	74732242	74732546
	ENST00000392485.2	17	exon 1	74732880	74733242
STAG2	ENST00000371160.1	X	exon 3	123156477	123156521
	ENST00000371160.1	X	exon 4	123159689	123159768
	ENST00000371160.1	X	exon 5	123164810	123164975
	ENST00000371160.1	X	exon 6	123171376	123171473
	ENST00000371160.1	X	exon 7	123176418	123176495
	ENST00000371160.1	X	exon 8	123179013	123179218
	ENST00000371160.1	X	exon 9	123181203	123181355
	ENST00000371160.1	X	exon 10	123182854	123182928
	ENST00000371160.1	X	exon 11	123184035	123184159
	ENST00000371160.1	X	exon 12	123184970	123185069
	ENST00000371160.1	X	exon 13	123185164	123185244
	ENST00000371160.1	X	exon 14	123189977	123190085
	ENST00000371160.1	X	exon 15	123191715	123191827
	ENST00000371160.1	X	exon 16	123195073	123195191
	ENST00000371160.1	X	exon 17	123195620	123195724
	ENST00000371160.1	X	exon 18	123196751	123196844
	ENST00000371160.1	X	exon 19	123196965	123197055
	ENST00000371160.1	X	exon 20	123197697	123197901
	ENST00000371160.1	X	exon 21	123199725	123199796
	ENST00000371160.1	X	exon 22	123200024	123200112
	ENST00000371160.1	X	exon 23	123200205	123200286
	ENST00000371160.1	X	exon 24	123202413	123202506
	ENST00000371160.1	X	exon 25	123204998	123205173
	ENST00000371160.1	X	exon 26	123210181	123210321
	ENST00000371160.1	X	exon 27	123211806	123211908

Gene	Transcript	Chromosome	Location	Start	End
STAG2 (cont.)	ENST00000371160.1	X	exon 28	123215229	123215378
	ENST00000371160.1	X	exon 29	123217270	123217399
	ENST00000371160.1	X	exon 30	123220396	123220620
	ENST00000371160.1	X	exon 31	123224424	123224614
	ENST00000371160.1	X	exon 33	123227867	123227994
	ENST00000371160.1	X	exon 34	123229221	123229299
	ENST00000371160.1	X	exon 35	123234423	123234447
TET2	ENST00000380013.4	4	exon 3	106155099	106158508
	ENST00000380013.4	4	exon 4	106162495	106162587
	ENST00000380013.4	4	exon 5	106163990	106164084
	ENST00000380013.4	4	exon 6	106164726	106164935
	ENST00000380013.4	4	exon 7	106180775	106180926
	ENST00000380013.4	4	exon 8	106182915	106183005
	ENST00000380013.4	4	exon 9	106190766	106190904
	ENST00000380013.4	4	exon 10	106193720	106194075
	ENST00000380013.4	4	exon 11	106196204	106197676
TP53	ENST00000269305.4	17	exon 11	7572925	7573008
	ENST00000269305.4	17	exon 10	7573926	7574033
	ENST00000269305.4	17	exon 9	7576852	7576926
	ENST00000269305.4	17	exon 8	7577018	7577155
	ENST00000269305.4	17	exon 7	7577498	7577608
	ENST00000269305.4	17	exon 6	7578176	7578289
	ENST00000269305.4	17	exon 5	7578370	7578554
	ENST00000269305.4	17	exon 4	7579311	7579590
	ENST00000269305.4	17	exon 3	7579699	7579721
	ENST00000269305.4	17	exon 2	7579838	7579912
U2AF1	ENST00000291552.4	21	exon 6	44514764	44514898
	ENST00000291552.4	21	exon 2	44524424	44524512
WT1	ENST00000332351.3	11	exon 1	32456245	32456892
	ENST00000332351.3	11	exon 2	32450042	32450165
	ENST00000332351.3	11	exon 3	32449501	32449604
	ENST00000332351.3	11	exon 4	32439122	32439200
	ENST00000332351.3	11	exon 5	32438035	32438086
	ENST00000332351.3	11	exon 6	32421493	32421590
	ENST00000332351.3	11	exon 7	32417802	32417953
	ENST00000332351.3	11	exon 8	32414211	32414301
	ENST00000332351.3	11	exon 9	32413517	32413610
	ENST00000332351.3	11	exon 10	32410603	32410725
ZRSR2	ENST00000307771.7	X	exon 1	15808617	15808659
	ENST00000307771.7	X	exon 2	15809056	15809136
	ENST00000307771.7	X	exon 3	15817994	15818076
	ENST00000307771.7	X	exon 4	15821810	15821919
	ENST00000307771.7	X	exon 5	15822233	15822320
	ENST00000307771.7	X	exon 6	15826355	15826394
	ENST00000307771.7	X	exon 7	15827322	15827441
	ENST00000307771.7	X	exon 8	15833799	15834013
	ENST00000307771.7	X	exon 9	15836709	15836765
	ENST00000307771.7	X	exon 10	15838329	15838439
	ENST00000307771.7	X	exon 11	15840853	15841383
rs10789158		1		63936188	63936308
rs3916765		6		32749936	32750056
rs1364429		7		134714538	134714658
rs2286510		17		9259404	9259524

Table C.3 ‘SNP backbone’ for mCA detection. *Supplementary material for Section 4.2.2.*

	number of targeted SNPs	**mean gap between SNPs (kb)**
chr1	811	305
chr2	877	277
chr3	735	269
chr4	683	280
chr5	665	272
chr6	639	267
chr7	559	285
chr8	524	279
chr9	422	334
chr10	491	276
chr11	483	279
chr12	475	282
chr13	367	262
chr14	341	256
chr15	298	276
chr16	280	322
chr17	283	287
chr18	291	268
chr19	196	300
chr20	227	277
chr21	127	303
chr22	122	279
chrX	430	355
	total = 10326 SNPs	**overall mean = 286 kb**

Table C.4 Regions targeted for detection of chromosomal rearrangements and KMT2A-PTD. The percentage of each region missing due to non-targetable highly repetitive sequences is shown. *Supplementary material for Section 4.2.2.*

Chromosomal rearrangement	Gene	Transcript	Chromosome	Location	Start	End	% missing due to repetitive regions
t(6;9); DEK-NUP214	DEK	ENST00000397239.3	6	intron 9-10	18226473	18236682	66
	NUP214	ENST00000359428.5	9	intron 17-18	134027281	134034769	46
t(8;21); RUNX1-RUNX1T1	RUNX1T1	ENST00000265814.3	8	intron 1-2	93029591	93088192	7
	RUNX1	ENST00000300305.3	21	intron 5-6	36206898	36231770	8
t(9;22); BCR-ABL	ABL1	ENST00000318560	9	intron 1-2	133710912	133729450	21
	ABL1	ENST00000318560	9	intron 2-3	133729624	133730187	0
	BCR	ENST00000305877.8	22	intron 1-2	23524426	23595985	14
	BCR	ENST00000305877.8	22	intron 13-14	23631808	23632525	0
	BCR	ENST00000305877.8	22	intron 14-15	23632600	23634727	17
	BCR	ENST00000305877.8	22	intron 19-20	23654023	23655073	0
t(15;17); PML-RARA	PML	ENST00000268058.3	15	intron 3-4	74315749	74317197	0
	PML	ENST00000268058.3	15	exon 6	74325496	74325755	0
	PML	ENST00000268058.3	15	intron 6-7	74325755	74326818	0
	RARA	ENST00000394089.2	17	intron 2-3	38487648	38504567	16
inv(16) or t(16;16); CBFB-MYH11	CBFB	ENST00000290858.6	16	intron 5-6	67116242	67132612	27
	CBFB	ENST00000290858.6	16	intron 4-5	67100701	67116115	29
	MYH11	ENST00000396324.3	16	intron 28-29	15820911	15826420	48
	MYH11	ENST00000396324.3	16	intron 29-30	15818849	15820704	12
	MYH11	ENST00000396324.3	16	intron 30-31	15818656	15818744	0
	MYH11	ENST00000396324.3	16	intron 31-32	15818266	15818503	0
	MYH11	ENST00000396324.3	16	intron 32-33	15815491	15818017	19
	MYH11	ENST00000396324.3	16	intron 33-34	15814908	15815278	0
	MYH11	ENST00000396324.3	16	intron 34-35	15814169	15814695	35
inv(3) or t(3;3); GATA2, MECOM			3	intergenic region	128294928	128324929	22
t(9;11); KMT2A-MLLT3	MLLT3	ENST00000380338.4	9	intron 5-6	20365742	20413718	7
	MLLT3	ENST00000380338.4	9	intron 8-9	20354877	20360739	4
	MLLT3	ENST00000380338.4	9	intron 4-5	20414423	20448120	5
	MLLT3	ENST00000380338.4	9	intron 9-10	20353594	20354805	0
	MLLT3	ENST00000380338.4	9	intron 6-7	20363603	20365666	11
	KMT2A	ENST00000534358.1	11	intron 7-8	118352807	118353136	0
	KMT2A	ENST00000534358.1	11	intron 8-9	118353210	118354897	56
	KMT2A	ENST00000534358.1	11	intron 9-10	118355029	118355576	0
Other KMT2A rearrangements	KMT2A	ENST00000534358.1	11	intron 10-11	118355690	118359328	29
	KMT2A	ENST00000534358.1	11	intron 11-12	118359475	118360506	23
KMT2A partial tandem duplication	KMT2A	ENST00000534358.1	11	exon 2	118339489	118339559	0
	KMT2A	ENST00000534358.1	11	exon 3	118342376	118345030	0
	KMT2A	ENST00000534358.1	11	exon 4	118347519	118347697	0
	KMT2A	ENST00000534358.1	11	exon 5	118348681	118348916	0
	KMT2A	ENST00000534358.1	11	exon 6	118350888	118350953	0
	KMT2A	ENST00000534358.1	11	exon 7	118352429	118352807	0
	KMT2A	ENST00000534358.1	11	exon 8	118353136	118353210	0
	KMT2A	ENST00000534358.1	11	exon 9	118354897	118355029	0
	KMT2A	ENST00000534358.1	11	exon 10	118355576	118355690	0
	KMT2A	ENST00000534358.1	11	exon 11	118359328	118359475	0
	KMT2A	ENST00000534358.1	11	exon 12	118360506	118360602	0
	KMT2A	ENST00000534358.1	11	exon 13	118360843	118360964	0
	KMT2A	ENST00000534358.1	11	exon 14	118361910	118362033	0
	KMT2A	ENST00000534358.1	11	exon 15	118362458	118362643	0
	KMT2A	ENST00000534358.1	11	exon 16	118363771	118363945	0
	KMT2A	ENST00000534358.1	11	exon 17	118365002	118365113	0
	KMT2A	ENST00000534358.1	11	exon 18	118365408	118365482	0
	KMT2A	ENST00000534358.1	11	exon 19	118366414	118366608	0
	KMT2A	ENST00000534358.1	11	exon 20	118366975	118367082	0
	KMT2A	ENST00000534358.1	11	exon 21	118368650	118368788	0
	KMT2A	ENST00000534358.1	11	exon 22	118369084	118369243	0
	KMT2A	ENST00000534358.1	11	exon 23	118370017	118370135	0
	KMT2A	ENST00000534358.1	11	exon 24	118370549	118370628	0
	KMT2A	ENST00000534358.1	11	exon 25	118371701	118371862	0
	KMT2A	ENST00000534358.1	11	exon 26	118372386	118372572	0
	KMT2A	ENST00000534358.1	11	exon 27	118373112	118377361	0

Table C.5 Variants present in Myeloid Reference Standard (Horizon Discovery Ltd) that should be detectable with our custom SNV/ indel panel. The reference standard was sequenced undiluted (100% reference standard), as well as diluted to 10%, 1% and 0.1% by mixing it with sample DNA from a 65 year old individual. *Supplementary material for Section 4.4*

			Expected VAF (%) in reference standard dilution			
Gene	**Variant**	**Variant type**	**undiluted**	**25% dilution**	**1% dilution**	**0.1% dilution**
ASXL1	W796C	SNV	5	1.25	0.05	0.005
BCOR	Q1174fs*8	Insertion	70	17.5	0.70	0.070
CBL	S403F	SNV	5	1.25	0.05	0.005
EZH2	R418Q	SNV	5	1.25	0.05	0.005
FLT3	D835Y	SNV	5	1.25	0.05	0.005
FLT3	Internal Tandem Duplication (ITD) (300 bp)		5	1.25	0.05	0.005
IDH1	R132C	SNV	5	1.25	0.05	0.005
IDH2	R172K	SNV	5	1.25	0.05	0.005
JAK2	F537-K539>L	Deletion	5	1.25	0.05	0.005
JAK2	V617F	SNV	5	1.25	0.05	0.005
KRAS	G13D	SNV	40	10.00	0.40	0.040
NPM1	W288fs*12	Insertion	5	1.25	0.05	0.005
NRAS	Q61L	SNV	10	2.50	0.10	0.010
RUNX1	M267I	SNV	35	8.75	0.35	0.035
SF3B1	G740E	SNV	5	1.25	0.05	0.005
TET2	R1261H	SNV	5	1.25	0.05	0.005
TP53	S241F	SNV	5	1.25	0.05	0.005

Table C.6 Variant calls for serial dilutions of Horizon Myeloid Reference Standard DNA from DCS and SSCS. Variants that were not detected are represented by '-'. **a.** Variant calls from DCS. **b.** Variant calls from SSCS. *Supplementary material for Section 4.4*

Expected VAF of variants: □ >10% □ 1-10% □ 0.1-1% □ 0.01-0.1% ■ 0.001-0.01%

a

DCS	100% Myeloid Reference Standard VAFs			25% Dilution VAFs			1% Dilution VAFs			0.1% Dilution VAFs		
Validated variant	Expected (%)	VAF1 (%)	VAF2 (%)	Expected (%)	VAF1 (%)	VAF2 (%)	Expected (%)	VAF1 (%)	VAF2 (%)	Expected (%)	VAF1 (%)	VAF2 (%)
BCOR p.Q1174fs*8	70	59.2	60.6	17.5	10.08	10.42	0.70	0.056	0.176	0.070	0.045	-
KRAS p.G13D	40	37.3	35.9	10.0	10.78	11.40	0.40	0.088	0.126	0.040	0.060	0.056
RUNX1 p.M267I	35	34.0	32.8	8.75	9.61	10.14	0.35	0.167	0.254	0.035	0.148	0.092
NRAS p.Q61L	10	8.6	9.1	2.50	2.88	2.70	0.10	0.081	-	0.010	-	-
ASXL1 p.W796C	5	4.8	4.7	1.25	1.67	1.39	0.05	-	0.038	0.005	0.040	-
CBL p.S403F	5	4.7	5.2	1.25	1.60	1.56	0.05	-	-	0.005	-	-
EZH2 p.R418Q	5	4.0	4.4	1.25	1.08	1.25	0.05	0.091	-	0.005	-	-
FLT3 ITD	5	4.1	3.5	1.25	1.11	1.15	0.05	-	-	0.005	-	-
FLT3 p.D835Y	5	4.4	4.4	1.25	1.40	1.16	0.05	-	-	0.005	-	-
IDH1 p.R132C	5	5.1	4.6	1.25	1.32	1.13	0.05	-	0.134	0.005	0.199	0.058
IDH2 p.R172K	5	4.4	5.0	1.25	1.44	1.34	0.05	0.054	0.098	0.005	-	-
JAK2 p.F537-K539>L	5	3.7	4.2	1.25	1.18	1.37	0.05	-	-	0.005	-	-
JAK2 p.V617F	5	4.4	5.8	1.25	0.92	1.28	0.05	-	-	0.005	-	-
NPM1 p.W288fs*12	5	4.2	3.9	1.25	1.21	1.03	0.05	-	-	0.005	-	-
SF3B1 p.G740E	5	4.6	4.8	1.25	1.12	1.63	0.05	-	0.169	0.005	-	-
TET2 p.R1261H	5	4.9	4.4	1.25	1.56	1.55	0.05	0.129	-	0.005	-	-
TP53 p.S241F	5	4.8	5.1	1.25	1.54	1.42	0.05	0.067	0.219	0.005	-	0.049
% detected in DCS		100%	100%		100%	100%		47.1%	47.1%		29.4%	23.5%

b

SSCS	100% Myeloid Reference Standard VAFs			25% Dilution VAFs			1% Dilution VAFs			0.1% Dilution VAFs		
Validated variant	Expected (%)	VAF1 (%)	VAF2 (%)	Expected (%)	VAF1 (%)	VAF2 (%)	Expected (%)	VAF1 (%)	VAF2 (%)	Expected (%)	VAF1 (%)	VAF2 (%)
BCOR p.Q1174fs*8	70	62.1	60.9	17.5	10.27	10.67	0.70	0.168	0.230	0.070	0.057	-
KRAS p.G13D	40	38.2	36.3	10.0	10.93	11.34	0.40	0.134	0.126	0.040	0.035	0.048
RUNX1 p.M267I	35	34.1	32.5	8.75	9.62	9.99	0.35	0.237	0.238	0.035	0.108	0.071
NRAS p.Q61L	10	8.8	9.0	2.50	2.71	2.72	0.10	0.104	-	0.010	-	-
ASXL1 p.W796C	5	4.8	4.8	1.25	1.56	1.49	0.05	-	0.027	0.005	0.041	0.012
CBL p.S403F	5	5.0	5.2	1.25	1.37	1.56	0.05	0.052	-	0.005	-	-
EZH2 p.R418Q	5	4.2	4.5	1.25	0.98	1.21	0.05	0.076	0.043	0.005	-	0.039
FLT3 ITD	5	3.8	3.5	1.25	0.93	1.21	0.05	-	-	0.005	-	-
FLT3 p.D835Y	5	4.5	4.6	1.25	1.36	1.29	0.05	-	0.033	0.005	-	-
IDH1 p.R132C	5	5.1	4.8	1.25	1.47	1.14	0.05	0.024	0.139	0.005	0.155	0.051
IDH2 p.R172K	5	4.2	5.1	1.25	1.37	1.32	0.05	0.034	0.081	0.005	-	0.014
JAK2 p.F537-K539>L	5	3.5	4.0	1.25	1.15	1.27	0.05	-	-	0.005	-	-
JAK2 p.V617F	5	4.6	5.1	1.25	1.06	1.22	0.05	-	0.024	0.005	-	-
NPM1 p.W288fs*12	5	4.5	4.6	1.25	1.20	1.13	0.05	-	0.023	0.005	-	-
SF3B1 p.G740E	5	4.5	5.1	1.25	1.26	1.63	0.05	0.064	0.132	0.005	0.024	0.019
TET2 p.R1261H	5	4.8	4.5	1.25	1.46	1.55	0.05	0.077	-	0.005	0.033	0.046
TP53 p.S241F	5	4.7	5.1	1.25	1.56	1.57	0.05	0.054	0.137	0.005	0.029	0.041
% detected in SSCS		100%	100%		100%	100%		64.7%	70.5%		47.0%	53.0%

C.1 Beta-binomial error model for SNV variant calling

C.1.1 Estimating the ε and δ parameters in the beta-binomial model

Supplementary material for Section 4.5.2.

To determine the error rate (ε) and beta-binomial dispersion (δ) parameters, for each position in the panel, two different methods were considered:

- 'Method 1' involved estimating ε as $\sum$variant reads/$\sum$depth across all the samples at the position and estimating δ parameter using the method-of-moments estimator (Box 4.5.2a, Section 4.5.2).
- 'Method 2' involved a maximum likelihood approach to estimate ε and δ, which minimised the negative log likelihood of the model (Box 4.5.2b, Section 4.5.2). δ was initialised using the method-of-moments estimator for δ, and ε was initialised as $\sum$variant reads/$\sum$depth. If the initialised δ was <0, then an initialisation of $\delta = 10^{-4}$ was used instead. A lower bound for δ was set to limit δ/ε to $>10^{-8}$, as below this level the distribution is definitely binomial.

To test the different methods, a custom Python script was written to generate a simulated dataset of 40 samples. These samples were each 'sequenced', at a depth of 1800X, using a 'panel' which targeted 2500 positions each with a different combination of position-specific error rate (ε) and beta-binomial dispersion parameter (δ) (Box C.1.1). For each 'position', ε and δ were inferred using each of the 2 methods to determine which method inferred δ and ε values closest to their true values.

Box C.1.1: Creating a simulated dataset containing positions with beta-binomially distributed error rates

- 50 different ε values were chosen, evenly log-spaced between 10^{-4} and 1.
- 50 different δ values were chosen, evenly log-spaced between 10^{-5} and 1000.
- For every combination of δ and ε (50 x 50 positions), with a total depth of 5000 in each sample at each position:
  - If $\delta\varepsilon > 10^{-7}$:
    - A range of variant reads was calculated, which would include 99.999% of the beta-binomial distribution:

      `k = np.arange(0, betabinom.ppf(0.0.99999, total depth, α, β))`
    - The probability of each possible number of variant reads (k) was calculated and then multiplied by the total number of samples, in order to calculate the expected number of samples with that number of variant reads:

      `int(number_samples*(betabinom.pmf(k, total depth, α, β)))`
  - If $\delta\varepsilon < 10^{-7}$ (i.e. binomial):
    - A range of variant reads was calculated, which would include 99.999% of the binomial distribution:

      `k = np.arange(0, binom.ppf(0.0.99999, total depth, ε))`
    - The probability of each possible number of variant reads (k) was calculated and then multiplied by the total number of samples, in order to calculate the expected number of samples with that number of variant reads:

      `int(number_samples*(binom.pmf(k, total depth, ε)))`

Both methods generally yielded similar estimates for ε and δ, over a range of values, although the method-of-moments estimator tended to overestimate δ when δ was $> 10^2$ (Figure C.1a, b). For both methods, when looking at the inferred $\delta\varepsilon$, a clear distinction can be seen between positions whose error rates are inferred to be beta-binomially distributed and those inferred are binomially distributed, with the beta-binomial positions clustering at $\delta\varepsilon > 10^{-5}$. Although both methods appeared to perform well, we chose to use the MLE method due to its better performance with high δ values.

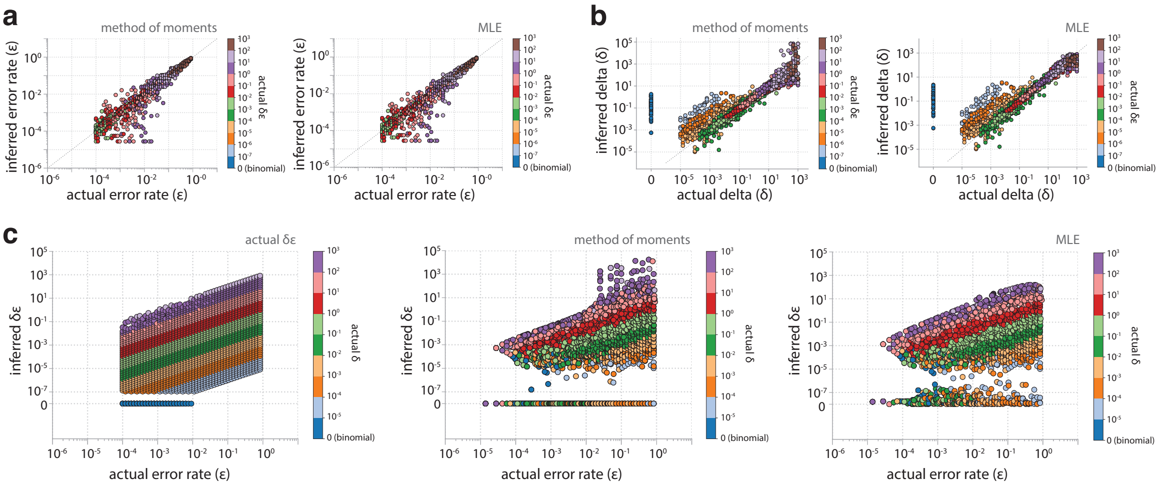


Figure C.1 Comparison of method-of-moments estimator and maximum likelihood approach (MLE) for inference of δ and ε using the simulated data (40 samples). a. Correlation between the actual position error rate (ε) and the ε inferred using $\sum$variant reads/$\sum$depth across all the samples at the position (left plot). Correlation between the actual ε and the ε inferred using the maximum likelihood approach (right plot). **b.** Correlation between the actual beta-binomial dispersion parameter (δ) and the δ inferred from the method-of-moments estimator (left plot). Correlation between the actual δ and the δ inferred using the maximum likelihood approach (right plot). Only positions where inferred $\delta > 0$ are shown. **c.** Left plot: The true values for ε vs $\delta\varepsilon$ in the simulated samples. Middle plot: The inferred values for ε vs $\delta\varepsilon$ using the method-of-moments approach. Right plot: The inferred values for ε vs $\delta\varepsilon$ using the maximum likelihood approach.

C.1.2 Relationship between ε and δ parameters in the beta-binomial model

Supplementary material for Section 4.5.2

The dispersion (δ) value that brings the beta-binomial distribution closer to binomial is different depending on the error rate (ε). The lower the ε, the higher the δ value required to bring the distribution closer to binomial. The parameter that seems to be important is $\delta\varepsilon$, which when $<10^{-6}$ results in the beta-binomial becoming binomial. When $\delta\varepsilon$ is $<10^{-13}$ the beta-binomial starts to 'fall apart' from the binomial, but this appears to be a numerical issue when α and β become very large numbers.

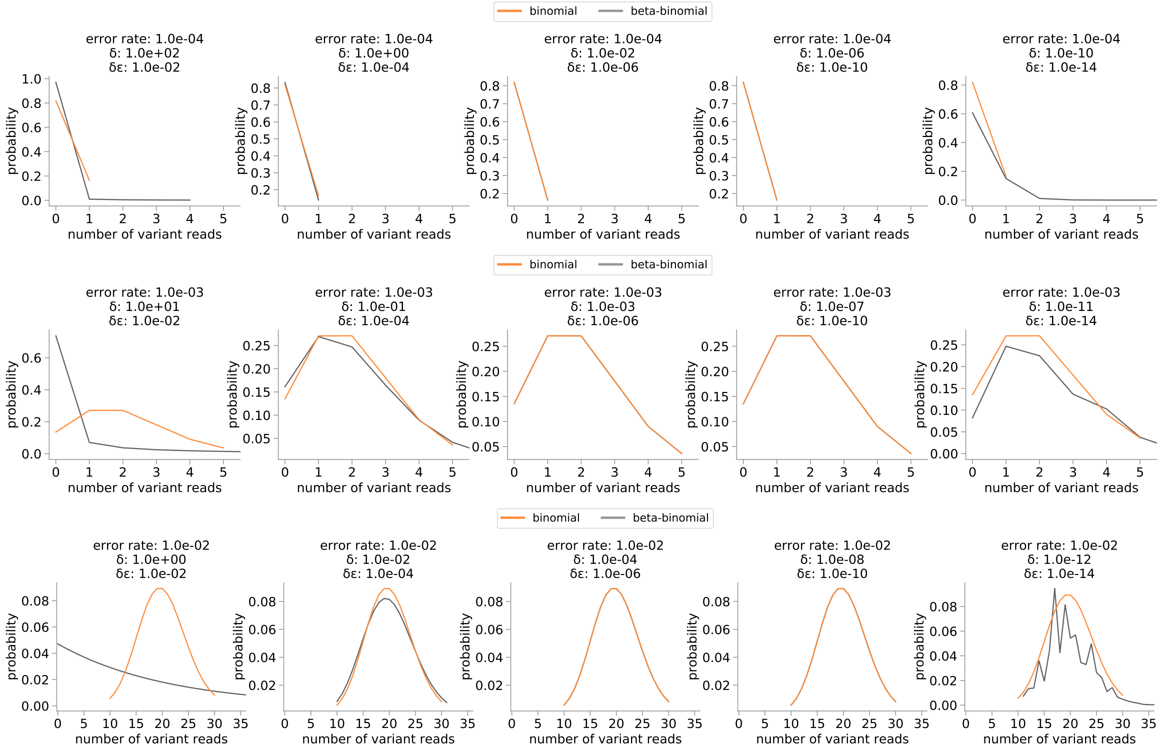


Figure C.2 Relationship between error rate and dispersion parameters. Simulated binomial and beta-binomial distributions are shown, for 3 different position-specific error rates (top row: 10^{-4}, middle row: 10^{-3}, bottom row: 10^{-2}) and different values for $\delta\varepsilon$ (columns: 10^{-2}, 10^{-4}, 10^{-6}, 10^{-10}, 10^{-14}).

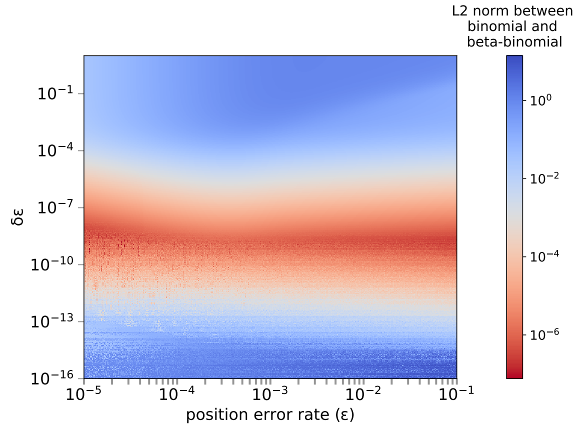


Figure C.3 $\delta\varepsilon$ values at which beta-binomial distribution collapses on to binomial. A mean depth of 2000 was used (different depths were also tried and heatmap looked effectively the same). For different combinations of ε and δ, the square difference between the beta-binomial likelihood ($f(k \mid N, \varepsilon, \delta)$ and binomial likelihood ($f(k \mid N, \varepsilon)$ was calculated and then summed across a range of variant read numbers (0.01-99.99% of the binomial and beta-binomial for that ε were calculated and the range of variant reads to sum across was taken from the minimum and maximum). The square root of the sum of these square differences was calculated (i.e. the L2 norm) and is represented by the colour on the plot, which shows ε vs $\delta\varepsilon$. Red areas represent ε and δ values where the beta-binomial distribution is closest to the binomial distribution (i.e. smallest L2 norm).

C.1.3 Approaches for determining the position-specific error distribution

Supplementary material for Section 4.5.2.

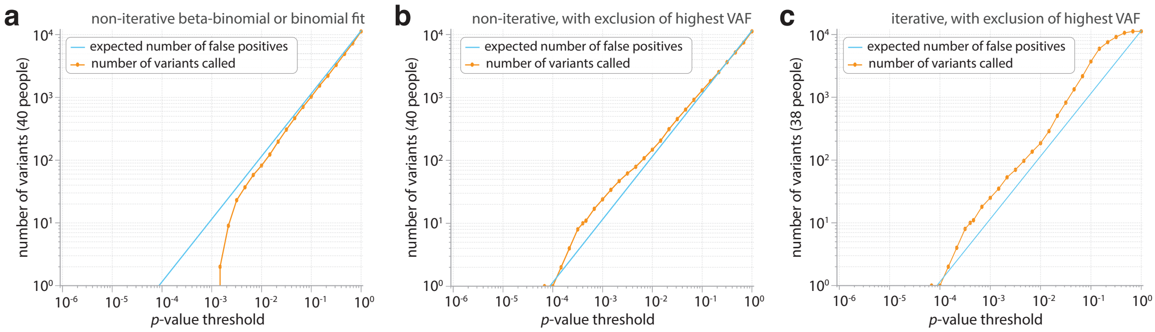


Figure C.4 Testing different approaches for determining the position-specific error distribution using simulated samples (all errors). a. A beta-binomial distribution (or binomial distribution if there were ≤ 3 samples with ≥ 1 variants reads) was fitted to all samples at the position. Samples were called as 'real' variants if their beta-binomial (or binomial) p-value was less than the p-value threshold. **b.** A beta-binomial distribution (or binomial distribution if there were ≤ 3 samples with ≥ 1 variants reads) was fitted to all samples at the position, except the sample with the highest VAF. Samples were called as 'real' variants if their beta-binomial (or binomial) p-value was less than the p-value threshold. **c.** An iterative approach was used, in which a beta-binomial distribution (or binomial distribution if there were ≤ 3 samples with ≥ 1 variants reads) was fitted to all samples at the position, except the sample with the highest VAF, variants were called as 'real' if their p-value was less than the p-value threshold; these variants were then excluded and the fitting process repeated. This was continued until no further 'real' variants were called at the position.

C.2 Developing a caller for chromosomal rearrangements

Supplementary material for Section 4.9.2.

C.2.1 Generating simulated samples containing chromosomal rearrangements

To understand the features we would expect to see in paired end sequencing data from samples containing a chromosomal rearrangement, as well as to test the performance of our caller, a custom Python script was written to generate simulated 'samples', each containing a different chromosomal rearrangement from each of the four chromosomal rearrangement categories (Figure 4.26). Simulated sequencing data was then generated from these samples, using a custom Python workflow that matched the steps of our custom panel workflow as much as possible.

t(9;22) BCR::ABL: Non-homologous chromosomes with relocated region in original orientation

To generate simulated samples containing t(9;22) BCR-ABL translocations, the full BCR and ABL1 sequences were downloaded from Ensembl (GRCh37 Release 104)[280]. To simulate the translocation, a random breakpoint position was chosen from the most commonly affected BCR and ABL1 breakpoint regions[166]: intron 1-2 of ABL1 and intron 14-15 of BCR (Figure C.5). BCR::ABL1 and ABL1::BCR fusion sequences were created by simply joining the ends of the sequences at the breakpoints together. The appropriate number of copies of BCR, ABL1, BCR-ABL1 fusion sequence and ABL1-BCR fusion sequence were created, according to the chosen number of simulated cells and VAF. For example, a simulated sample of 100 cells harbouring a BCR-ABL1 translocation at 25% VAF would contain 150 copies of normal BCR, 150 copies of normal ABL1, 50 copies of BCR-ABL1 and 50 copies of ABL1::BCR.

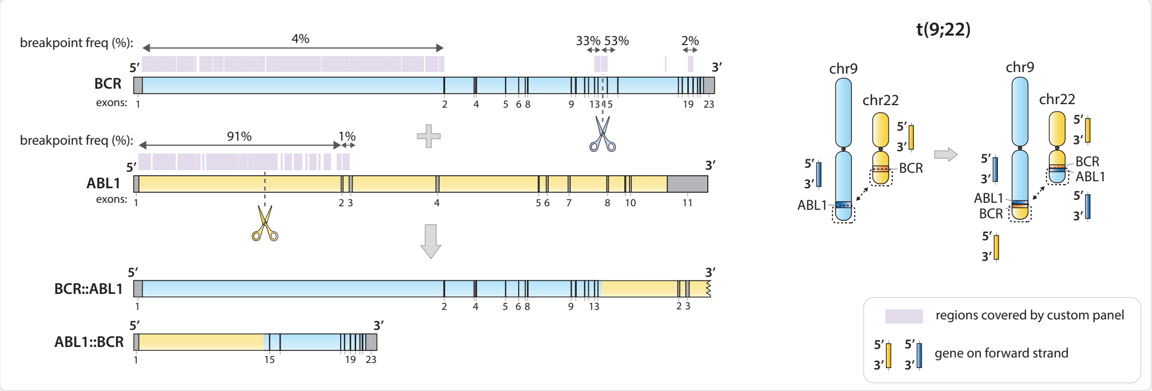


Figure C.5 Generating simulated samples containing t(9;22) BCR::ABL

t(3;3) GATA2::MECOM: Homologous chromosomes with relocated region in original orientation

The baits for targeting the breakpoint regions involved in t(3;3) in our custom panel were unfortunately designed for the wrong region of chromosome 3, and so it will not be possible to detect t(3;3) with the current version of our panel. Nonetheless, we wanted to be able to test if our caller could detect rearrangements in this category (homologous chromosomes with relocated region in original orientation), so when we generated simulated samples containing t(3;3) GATA2::MECOM rearrangements, we chose random breakpoint positions from within the region on chromosome 3 that our panel targeted (3' breakpoint between chr3 128294928-128295400 and 5' breakpoint between chr3 128324400-128324929) (Figure C.6a). Although both t(3;3) and inv(3) share a common 3' breakpoint region and both result in the relocation of the G2DHE region to near the MECOM region, the 5' breakpoint is different between the two rearrangements (Figure C.6). We focused on t(3;3) rather than inv(3) for our simulated sample because it was the only chromosomal rearrangement we targeted with our panel that was in the 'homologous chromosomes with relocated region in original orientation' class. Fusion sequences were created by simply joining the ends of the sequences at the breakpoints together. The appropriate number of copies were chosen according to the chosen number of simulated cells and VAF.

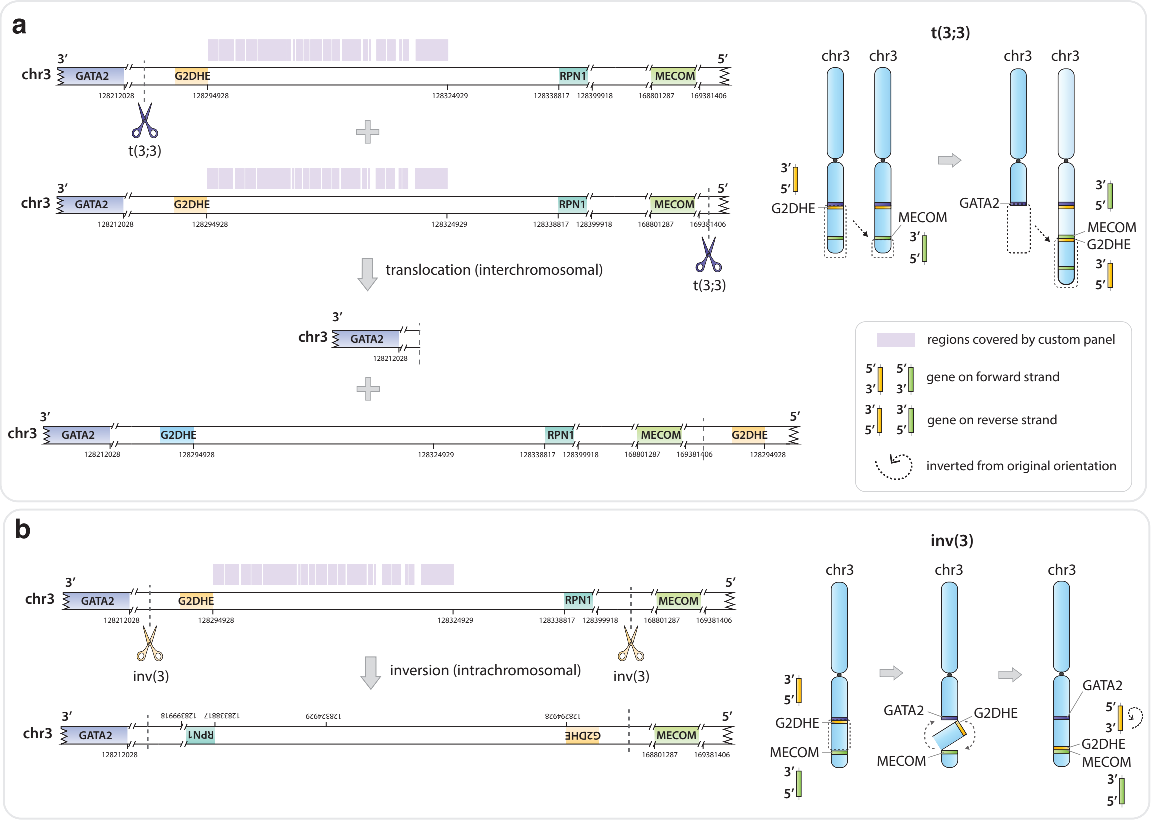


Figure C.6 Differences between t(3;3) and inv(3) a. t(3;3) GATA2::MECOM b. inv(3) GATA2::MECOM

t(9;11) KMT2A::MLLT3: Non-homologous chromosomes with relocated region inverted

To generate simulated samples containing t(9;11) KMT2A::MLLT3 translocations, the full KMT2A and MLLT3 sequences were downloaded from Ensembl (GRCh37 Release 104)[280]. To simulate the translocation, a random breakpoint position was chosen from the most commonly affected KMT2A and MLLT3 breakpoint regions[166]: intron 7-8 of KMT2A and intron 5-6 of MLLT3 (Figure C.7). KMT2A is located on the forward strand of the q arm of chromosome 11 and MLLT3 is located on the reverse strand of the p arm of chromosome 9. The t(9;11) KMT2A::MLLT3 translocation involves relocation of the end part of the MLLT3 gene to join next to the end of the first part of the KMT2A gene, with inversion of the relocated MLLT3 segment in the process. Similarly, the reciprocal translocation involves relocation of the end part of the KMT2A gene to join next to the end of the first part of the MLLT3 gene, with inversion of the relocated KMT2A segment in the process. If viewing from the perspective of the forward strand, the inverted segment is located on the 3' side of the KMT2A::MLLT3 fusion gene on the q arm of chromosome 11 and is located on the 5' side of the MLLT3::KMT2A fusion gene on the p arm chromosome 9. KMT2A::MLLT3 and MLLT3::KMT2A fusion sequences were created to reflect this and simulated samples were created, each containing 100 simulated cells with the translocation at a chosen VAF.

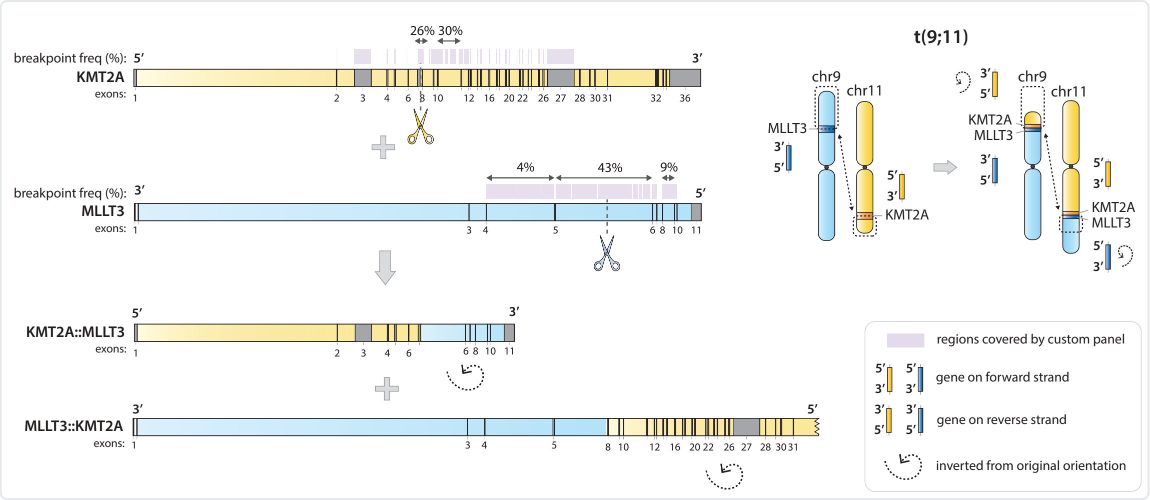

26% 30%
breakpoint freq (%): 5′ 3′
KMT2A
exons: 1 2 3 4 6 8 10 12 16 20 22 26 27 28 30 31 32 36
4% 43% 9%
breakpoint freq (%): 3′ 5′
MLLT3
exons: 1 3 4 5 6 8 10 11
5′ 3′
KMT2A::MLLT3
exons: 1 2 3 4 6 6 8 10 11
3′ 5′
MLLT3::KMT2A
exons: 1 3 4 5 8 10 12 16 20 22 26 27 28 30 31
t(9;11)
chr9
chr11
MLLT3
3′
5′
5′
3′
KMT2A
3′
5′
chr9
chr11
KMT2A
MLLT3
3′
5′
5′
3′
KMT2A
MLLT3
5′
3′
regions covered by custom panel
5′ 3′ 5′ 3′ gene on forward strand
3′ 5′ 3′ 5′ gene on reverse strand
inverted from original orientation

Figure C.7 Generating simulated samples containing t(9;11) KMT2A::MLLT3.

t(16;16) and inv(16) CBFB::MYH11: Homologous chromosomes with relocated region inverted

To generate simulated samples containing t(16;16) CBFB::MY11 and samples containing inv(16) CBFB::MYH11, the full CBFB and MYH11 sequences were downloaded from Ensembl (GRCh37 Release 104)[280]. To simulate the translocation or inversion, a random breakpoint position was chosen from the most commonly affected CBFB and MYH11 breakpoint regions[175]: intron 5-6 of CBFB and intron 33-34 of MYH11 (Figure C.8). In both t(16;16) and inv(16), the relocated gene regions are inverted compared to their original orientation. In t(16;16) it is because segments from the p and q arms interchange with each other (Figure C.8a), whereas in inv(16) it is because an inversion within the chromosome occurs (Figure C.8b). Both t(16;16) and inv(16) result in a CBFB::MYH11 and MYH11:CBFB fusion genes. In t(16;16), if viewing from the perspective of the forward strand, the inverted segment is located on the 3' side of the CBFB::MYH11 fusion gene on the q arm and the inverted segment is located on the 5' side of the MYH11::CBFB fusion gene on the p arm. In contrast, in inv(16), the inverted segment is located on the 3' side of the CBFB::MYH11 fusion gene on the p arm and on the 5' side of the MYH11::CBFB fusion gene on the q arm. CBFB::MYH11 and MYH11::CBFB fusion sequences, for both t(16;16) and inv(16) were created to reflect this and simulated samples were created, each containing 100 simulated cells with the chromosomal rearrangement at a chosen VAF.

C.2.2 Generating simulated sequencing data from the simulated samples

The sequences from the simulated samples were then 'fragmented', producing DNA fragment sizes normally distributed about 200 bp, which matches the size distribution obtained in our actual custom panel library preparation workflow. The sequences of our actual custom panel probes were then used to 'capture' the simulated DNA fragments if at least 30 consecutive nucleotides matched the sequence of the probe. An

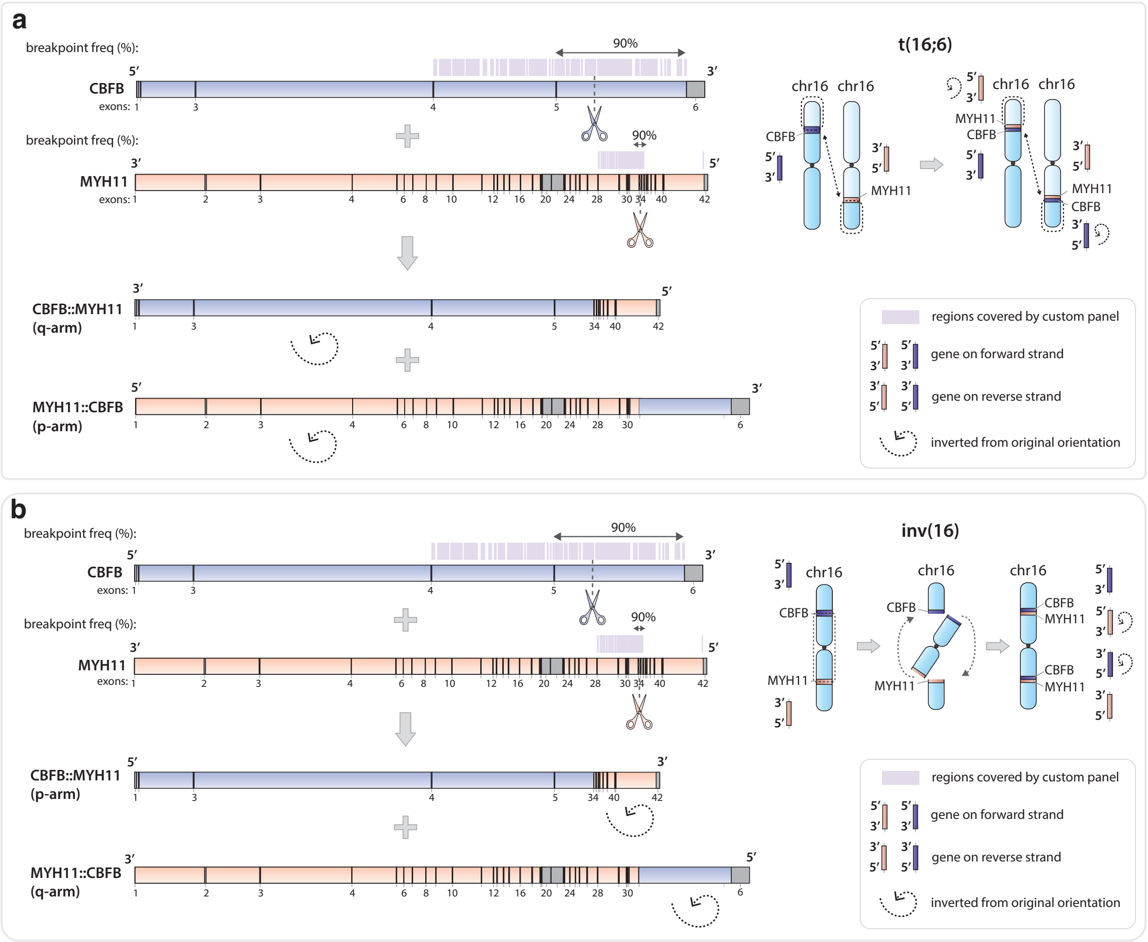


Figure C.8 Generating simulated samples containing t(16;16) or inv(16). a. t(16;16) CBFB::MYH11 b. inv(16) CBFB::MYH11

interleaved fastq file was then produced, to simulate paired end 150 bp ‘sequencing’ of the fragments, containing read1 and read2 sequences from both the forward and reverse reads, with 3bp UMI (duplex) at the start of each read. If the DNA fragment was <150 bp in length, then Illumina adapter sequences were appended to the end of the read, so that the total length of the read was 150 bp. The fastq files were then processed using the first part of our computational workflow to produce a mapped SSCS BAM file (Section 4.3.1). Illumina adapter sequences were hard-clipped and their position information stored in the ‘XT’ tag of the BAM file. UMI information was stored in the ‘RX’ tag of the BAM file. The mapped SSCS BAM files were used for exploring the types of reads present in each of the different four classes of chromosomal rearrangement and for testing our custom caller.

D

Supplementary material for Chapter 5

D.1 Clonal trajectories for pre-AML cases and matched controls

Supplementary material for Section 5.2

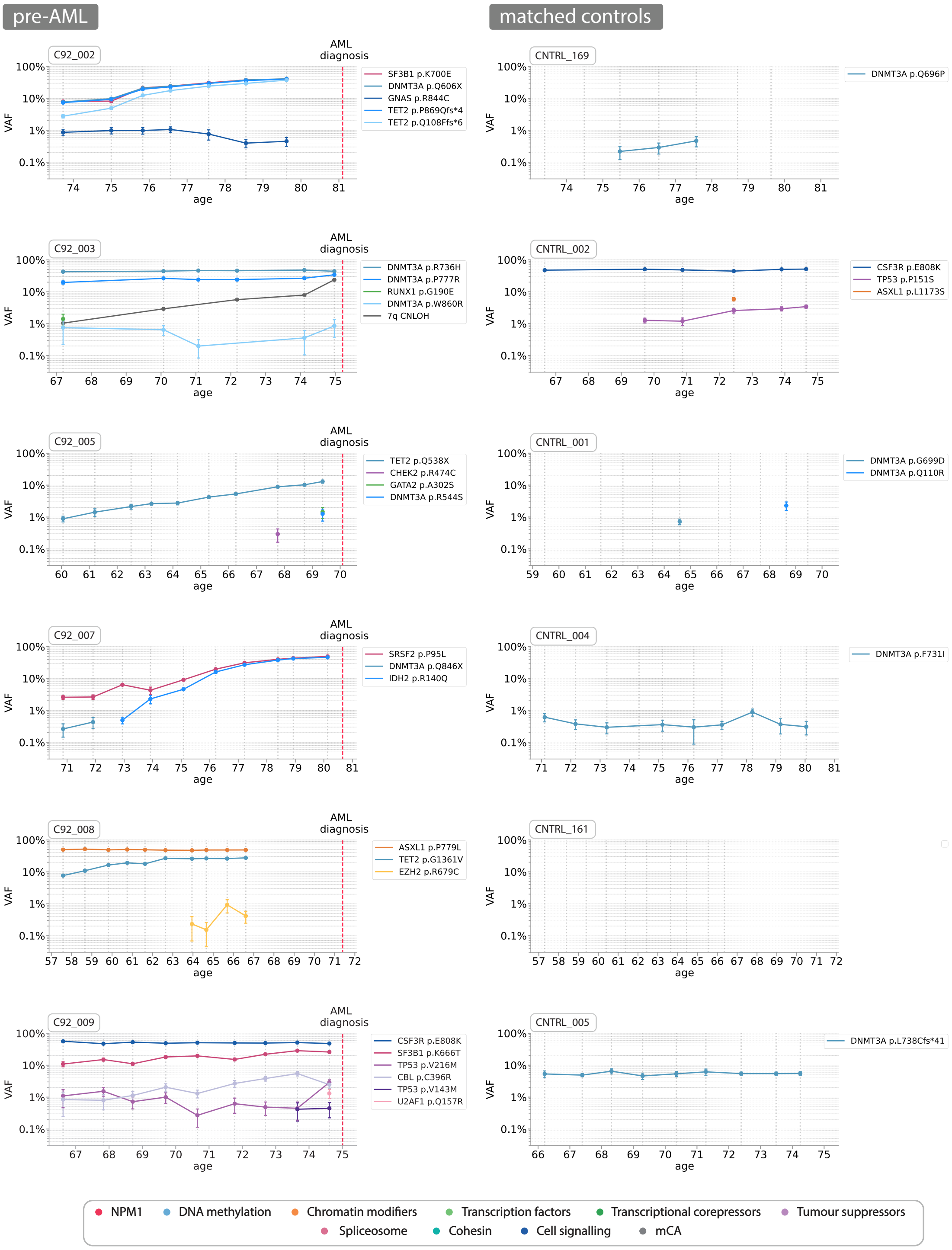


Figure D.1 Clonal trajectories for pre-AML cases and matched controls: Part 1. Grey vertical lines indicate blood sample timepoints. Error bars represent sampling error, taking in to account logistic growth of clones $(\pm\sqrt{(\text{variant depth} \times (1 - \text{VAF}))}/\text{position depth})$. Trajectories are coloured according to their class gene class (see legend).

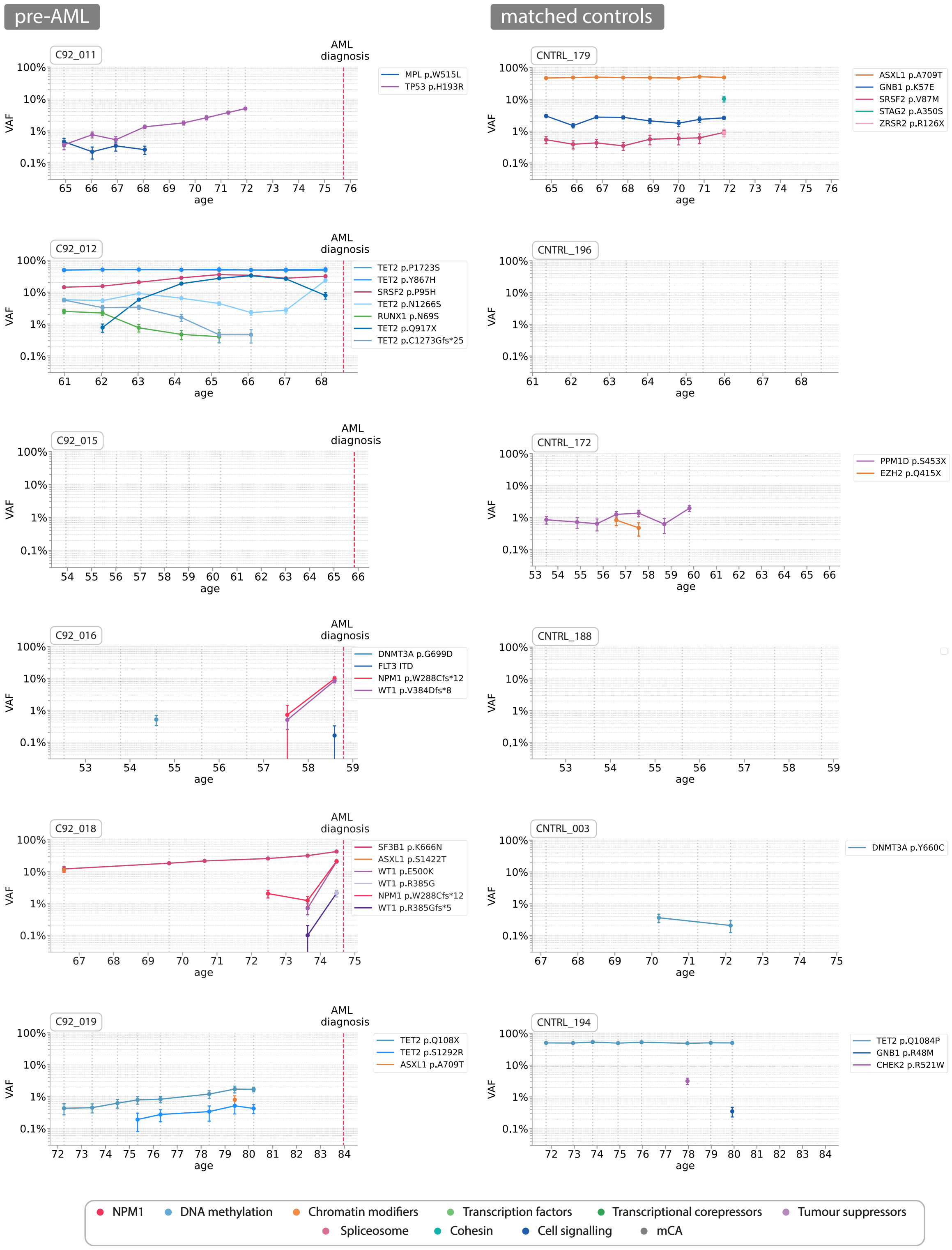


Figure D.2 **Clonal trajectories for pre-AML cases and matched controls: Part 2.** Grey vertical lines indicate blood sample timepoints. Error bars represent sampling error, taking in to account logistic growth of clones $(\pm\sqrt{(\text{variant depth} \times (1 - \text{VAF}))}/\text{position depth})$. Trajectories are coloured according to their class gene class (see legend).

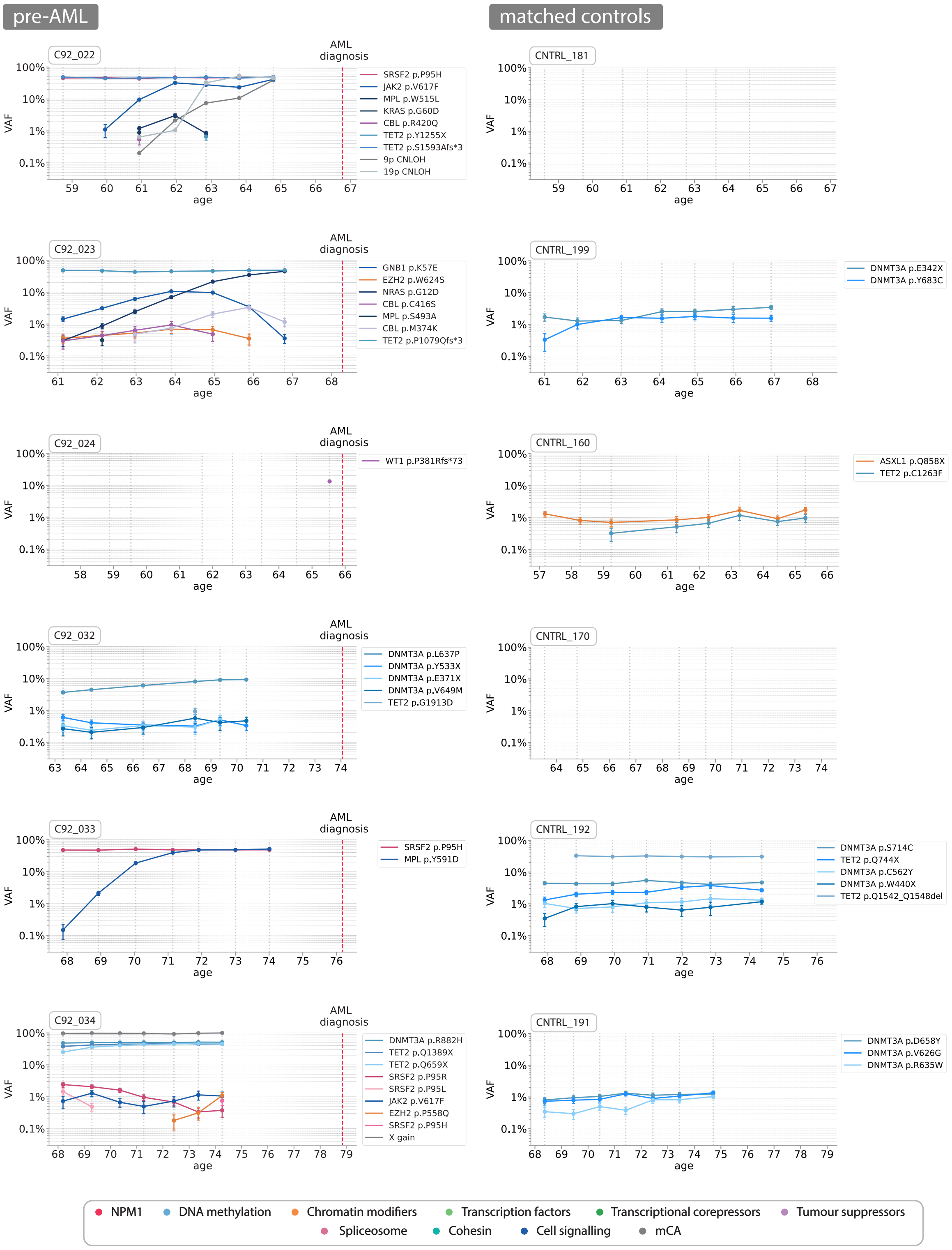


Figure D.3 Clonal trajectories for pre-AML cases and matched controls: Part 3. Grey vertical lines indicate the blood sample timepoints. Error bars are $\pm\sqrt{(\text{variant depth} \times (1-\text{VAF}))}/\text{position depth}$. Datapoints and trajectories are coloured according to their class of mutation (see legend).

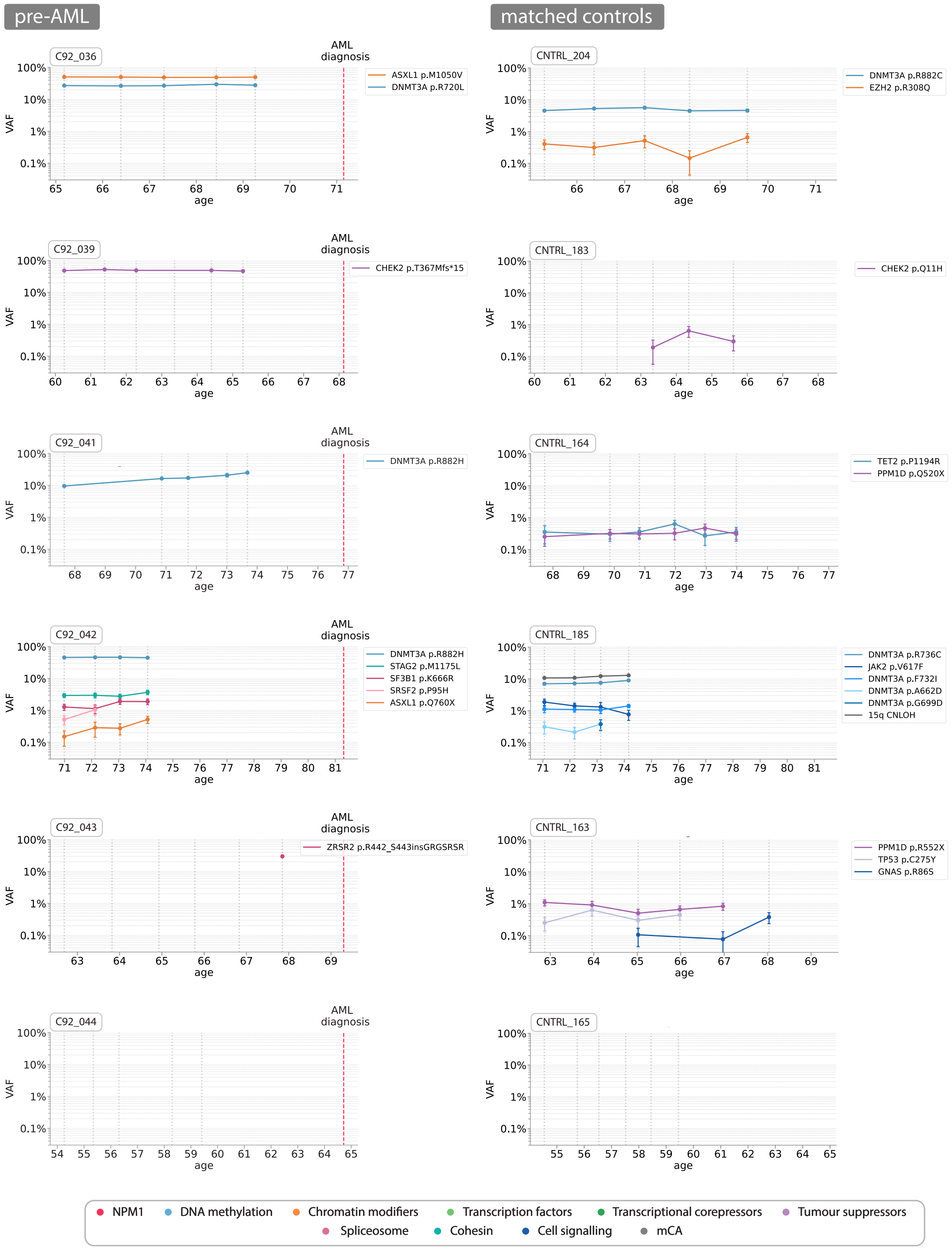


Figure D.4 **Clonal trajectories for pre-AML cases and matched controls: Part 4.** Grey vertical lines indicate blood sample timepoints. Error bars represent sampling error, taking in to account logistic growth of clones $(\pm\sqrt{(\text{variant depth} \times (1-\text{VAF}))}/\text{position depth})$. Trajectories are coloured according to their class gene class (see legend).

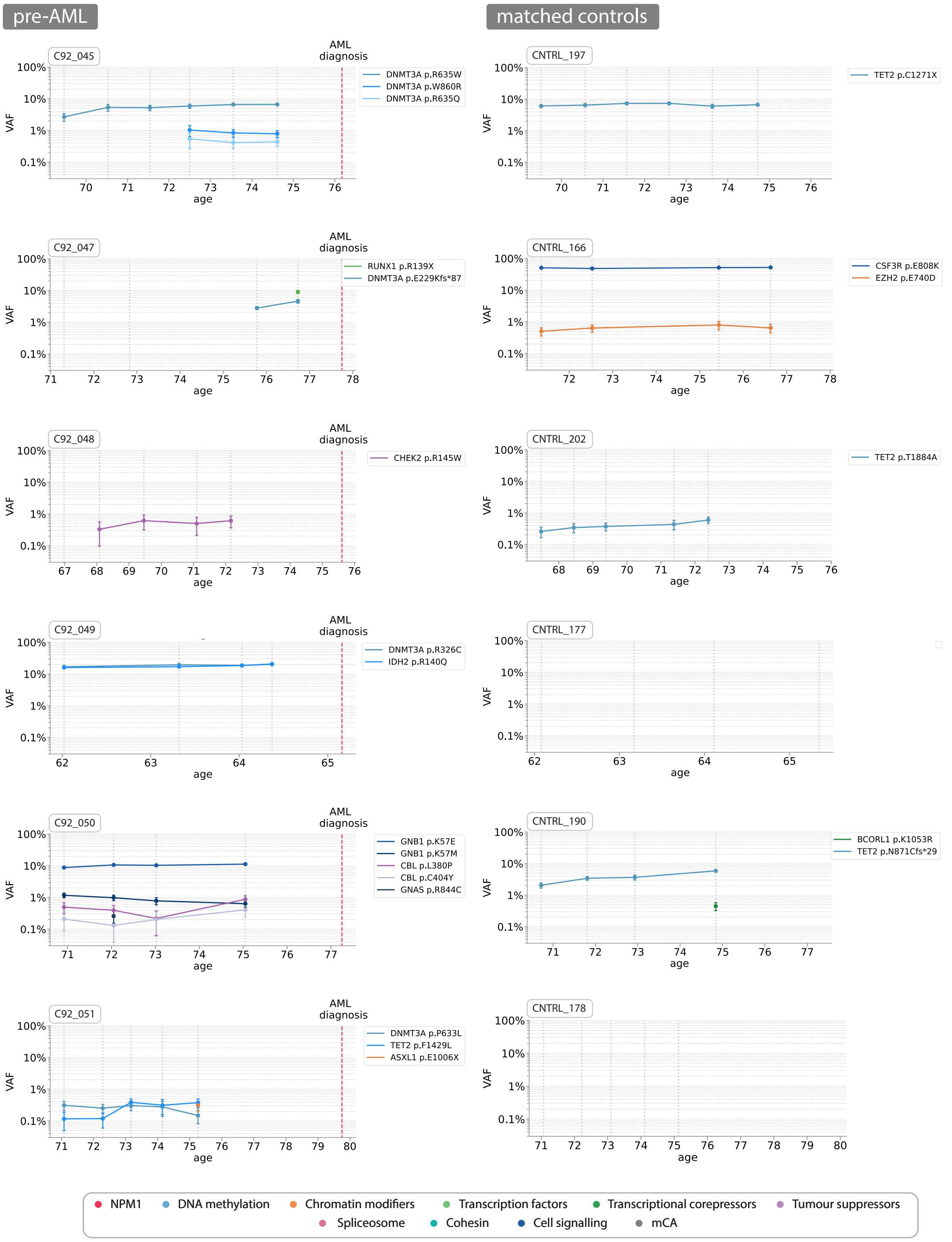


Figure D.5 **Clonal trajectories for pre-AML cases and matched controls: Part 5.** Grey vertical lines indicate blood sample timepoints. Error bars represent sampling error, taking in to account logistic growth of clones ($\pm\sqrt{(\text{variant depth} \times (1-\text{VAF}))}/\text{position depth}$). Trajectories are coloured according to their class gene class (see legend).

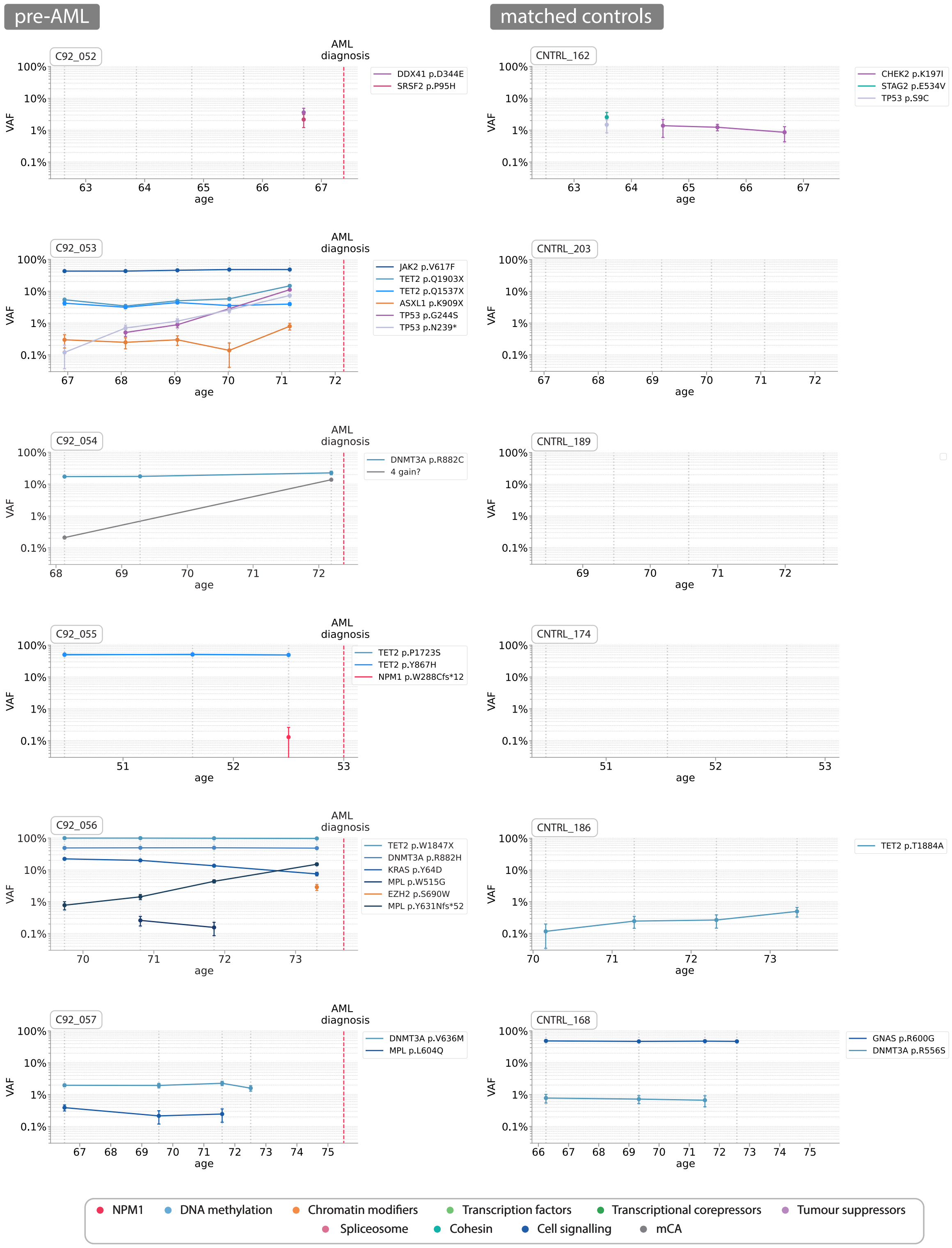


Figure D.6 Clonal trajectories for pre-AML cases and matched controls: Part 6. Grey vertical lines indicate blood sample timepoints. Error bars represent sampling error, taking in to account logistic growth of clones ($\pm\sqrt{(\text{variant depth} \times (1-\text{VAF}))}/\text{position depth}$). Trajectories are coloured according to their class gene class (see legend).

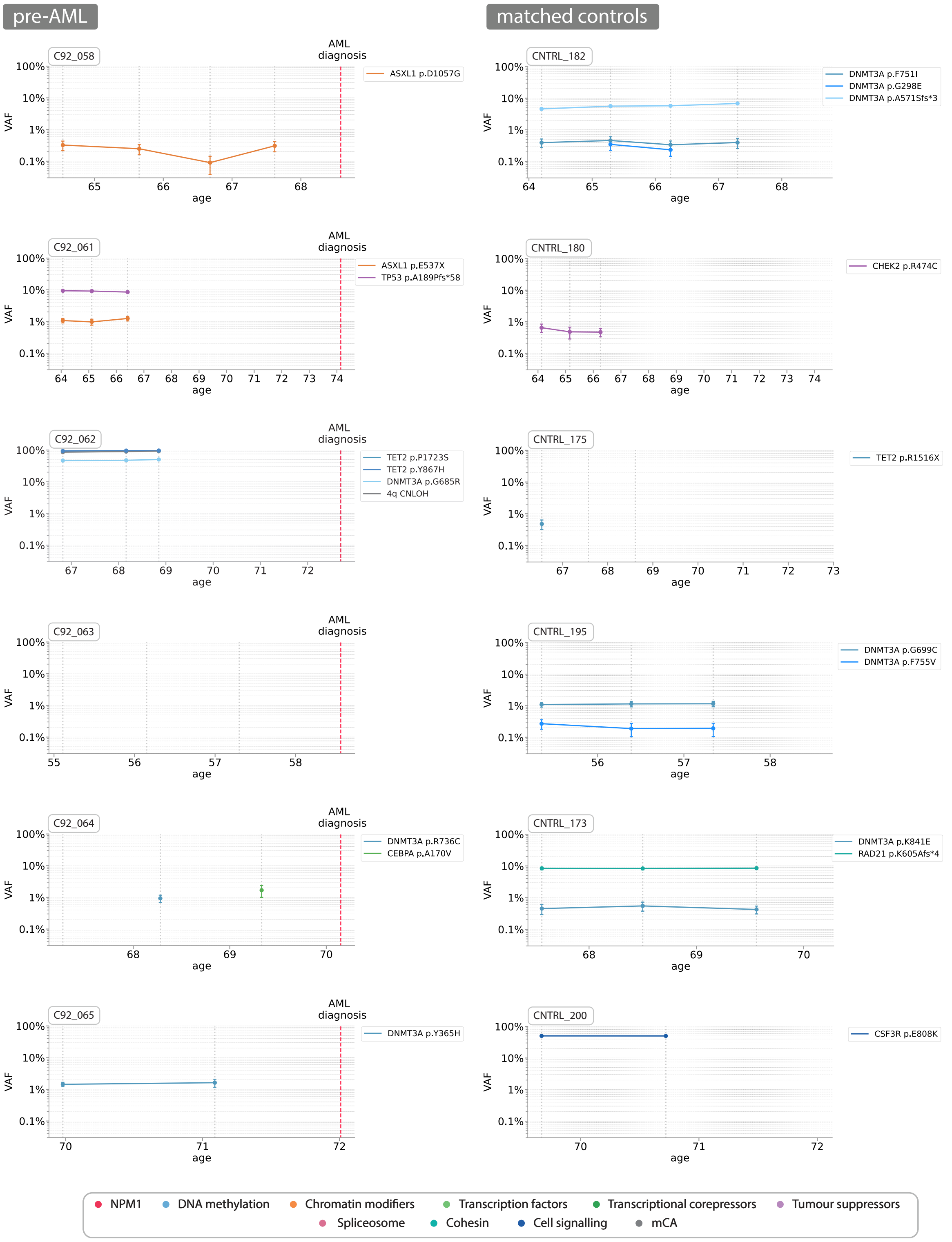


Figure D.7 Clonal trajectories for pre-AML cases and matched controls: Part 7. Grey vertical lines indicate blood sample timepoints. Error bars represent sampling error, taking in to account logistic growth of clones $(\pm\sqrt{(\text{variant depth} \times (1-\text{VAF}))}/\text{position depth})$. Trajectories are coloured according to their class gene class (see legend).

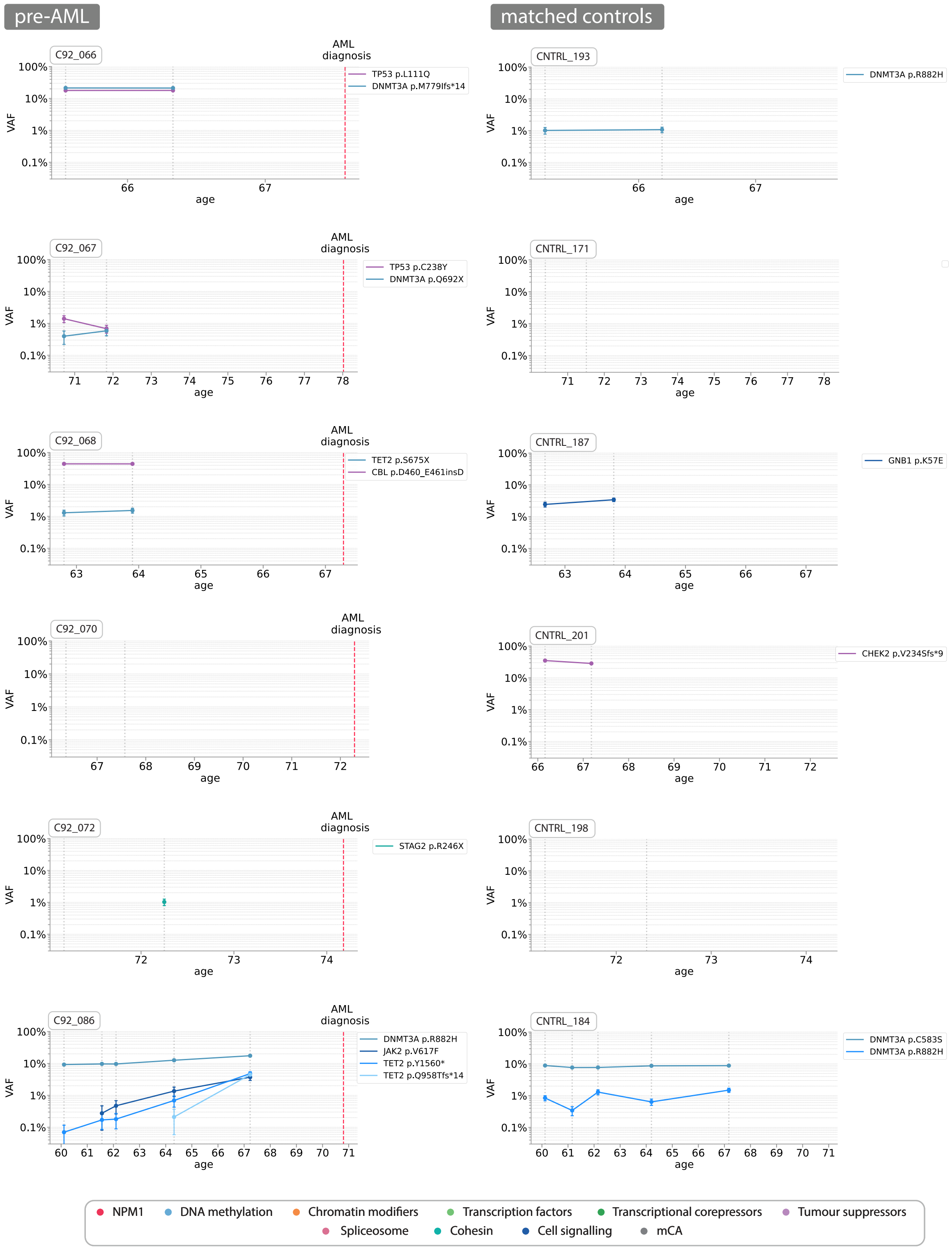


Figure D.8 Clonal trajectories for pre-AML cases and matched controls: Part 8. Grey vertical lines indicate blood sample timepoints. Error bars represent sampling error, taking in to account logistic growth of clones $(\pm\sqrt{(\text{variant depth} \times (1-\text{VAF}))}/\text{position depth})$. Trajectories are coloured according to their class gene class (see legend).

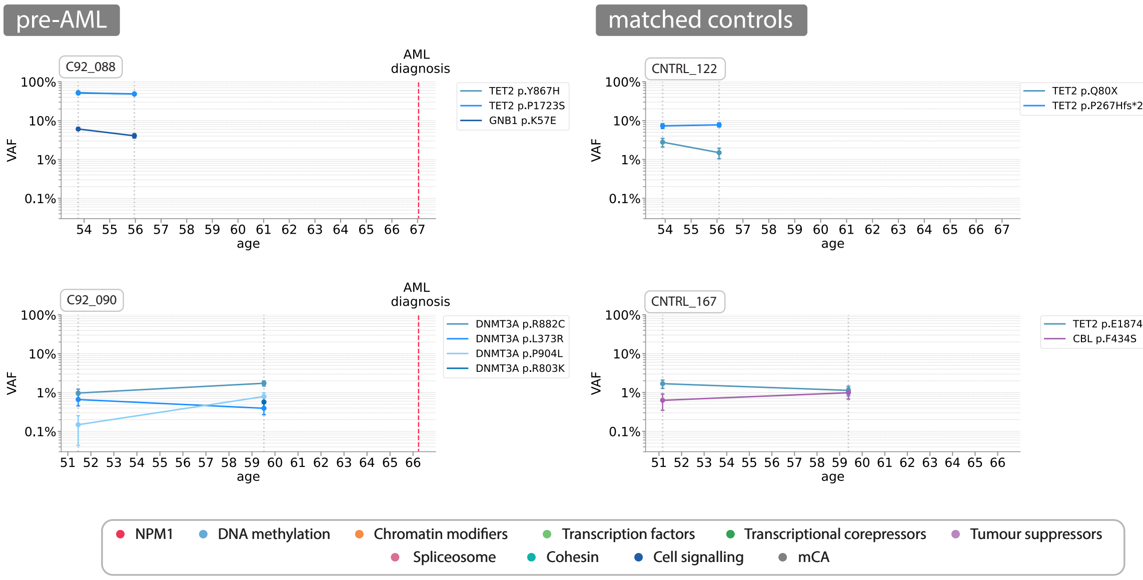


Figure D.9 Clonal trajectories for pre-AML cases and matched controls: Part 9. Grey vertical lines indicate blood sample timepoints. Error bars represent sampling error, taking in to account logistic growth of clones $(\pm\sqrt{(\text{variant depth} \times (1 - \text{VAF}))}/\text{position depth})$. Trajectories are coloured according to their class gene class (see legend).

D.2 Mutation acquisition age and fitness for linear evolution samples

Supplementary material for Section 5.3.1

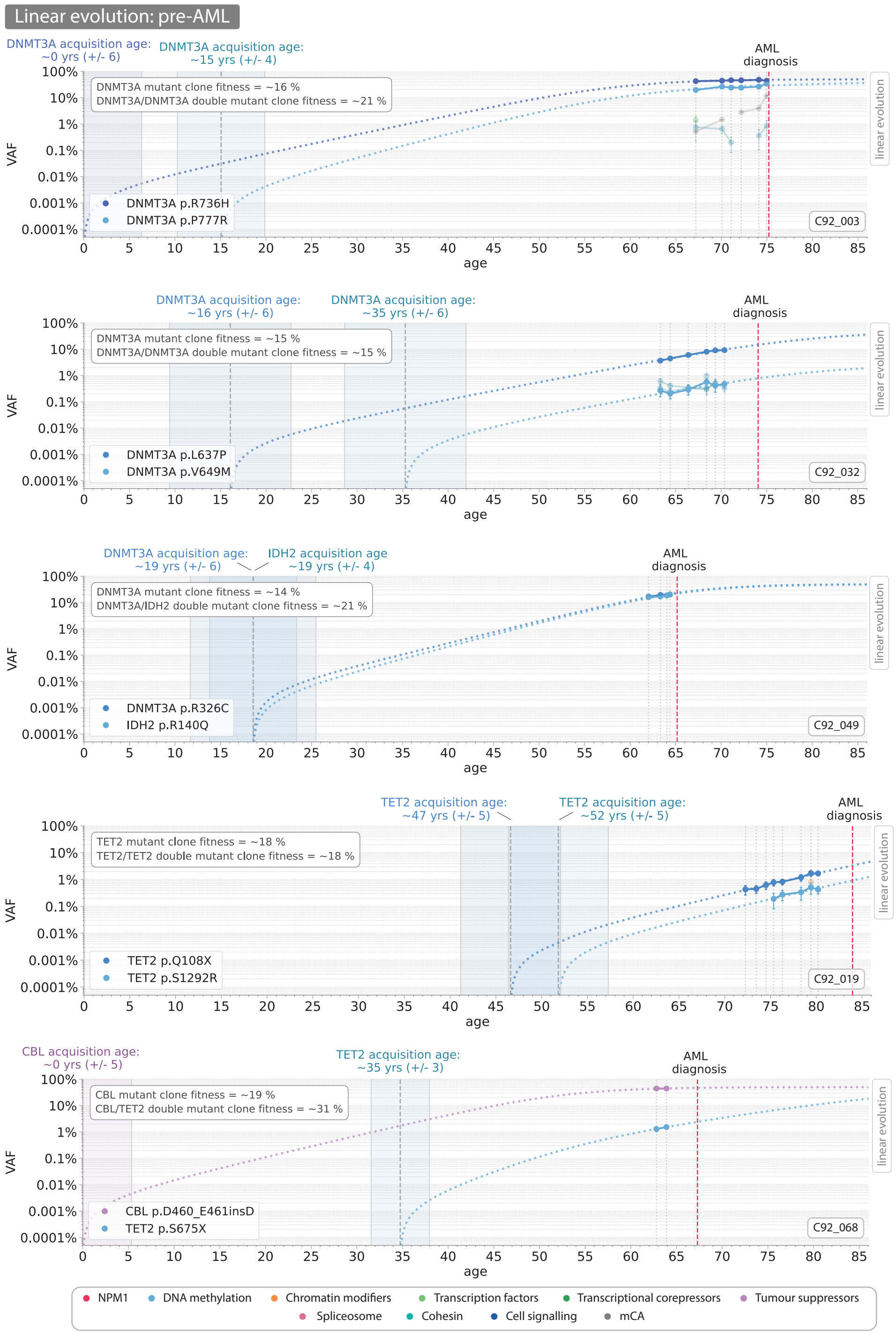


Figure D.10 Estimation of acquisition age and fitness for mutations showing a linear evolution pattern: Part 1. Grey vertical lines indicate blood sample timepoints. Trajectories are coloured according to their class gene class (see legend). Dash coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches. The error measurement shown for the acquisition ages is $\pm 1/s$. Other mutations present in the sample are shown as faded datapoints.

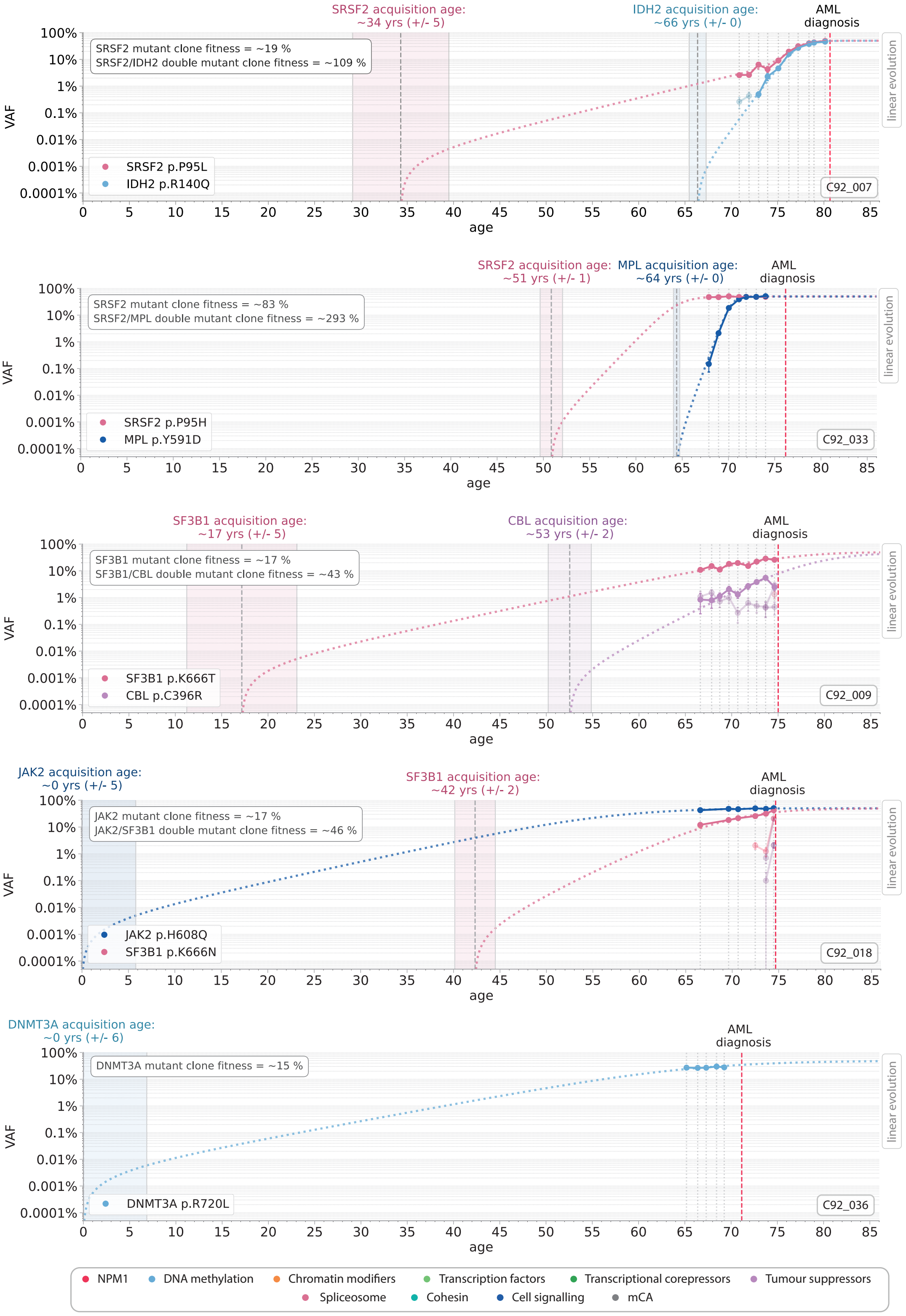


Figure D.11 Estimation of acquisition age and fitness for mutations showing a linear evolution pattern: Part 2. Grey vertical lines indicate blood sample timepoints. Trajectories are coloured according to their class gene class (see legend). Dash coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches. The error measurement shown for the acquisition ages is $\pm 1/s$. Other mutations present in the sample are shown as faded datapoints.

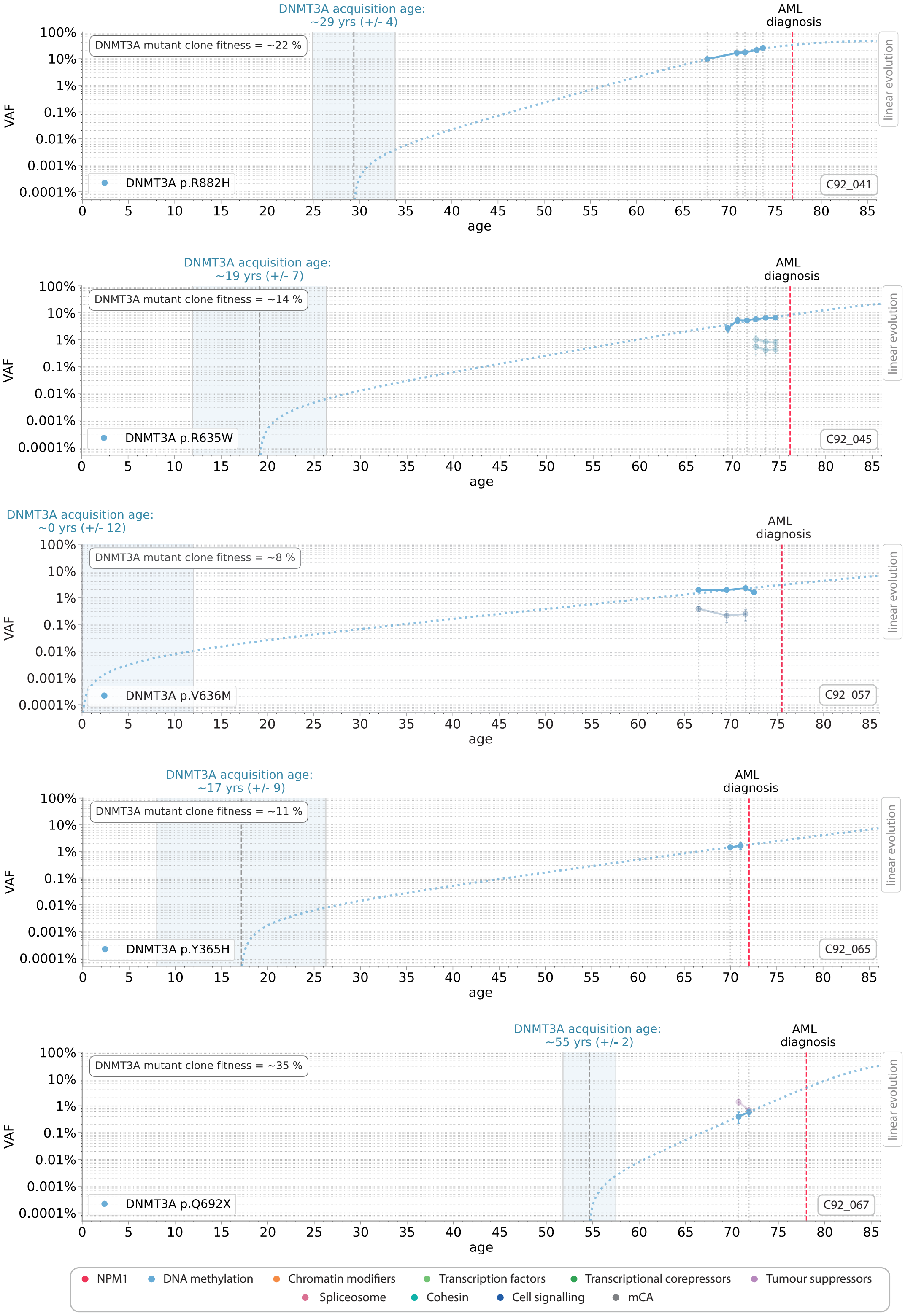


Figure D.12 **Estimation of acquisition age and fitness for mutations showing a linear evolution pattern: Part 3.** Grey vertical lines indicate blood sample timepoints. Trajectories are coloured according to their class gene class (see legend). Dash coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches. The error measurement shown for the acquisition ages is $\pm 1/s$. Other mutations present in the sample are shown as faded datapoints.

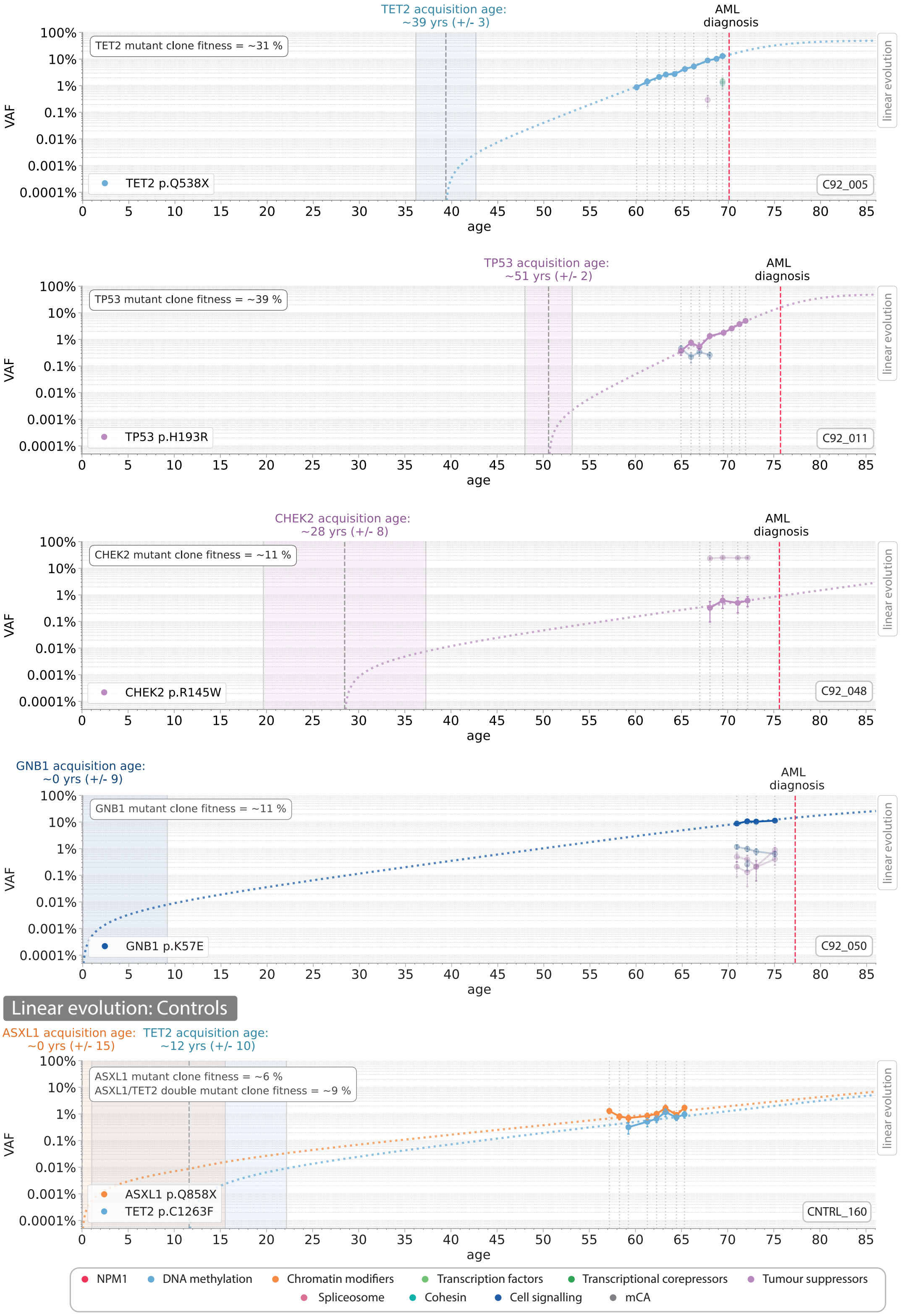


Figure D.13 Estimation of acquisition age and fitness for mutations showing a linear evolution pattern: Part 4. Grey vertical lines indicate blood sample timepoints. Trajectories are coloured according to their class gene class (see legend). Dash coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches. The error measurement shown for the acquisition ages is $\pm 1/s$. Other mutations present in the sample are shown as faded datapoints.

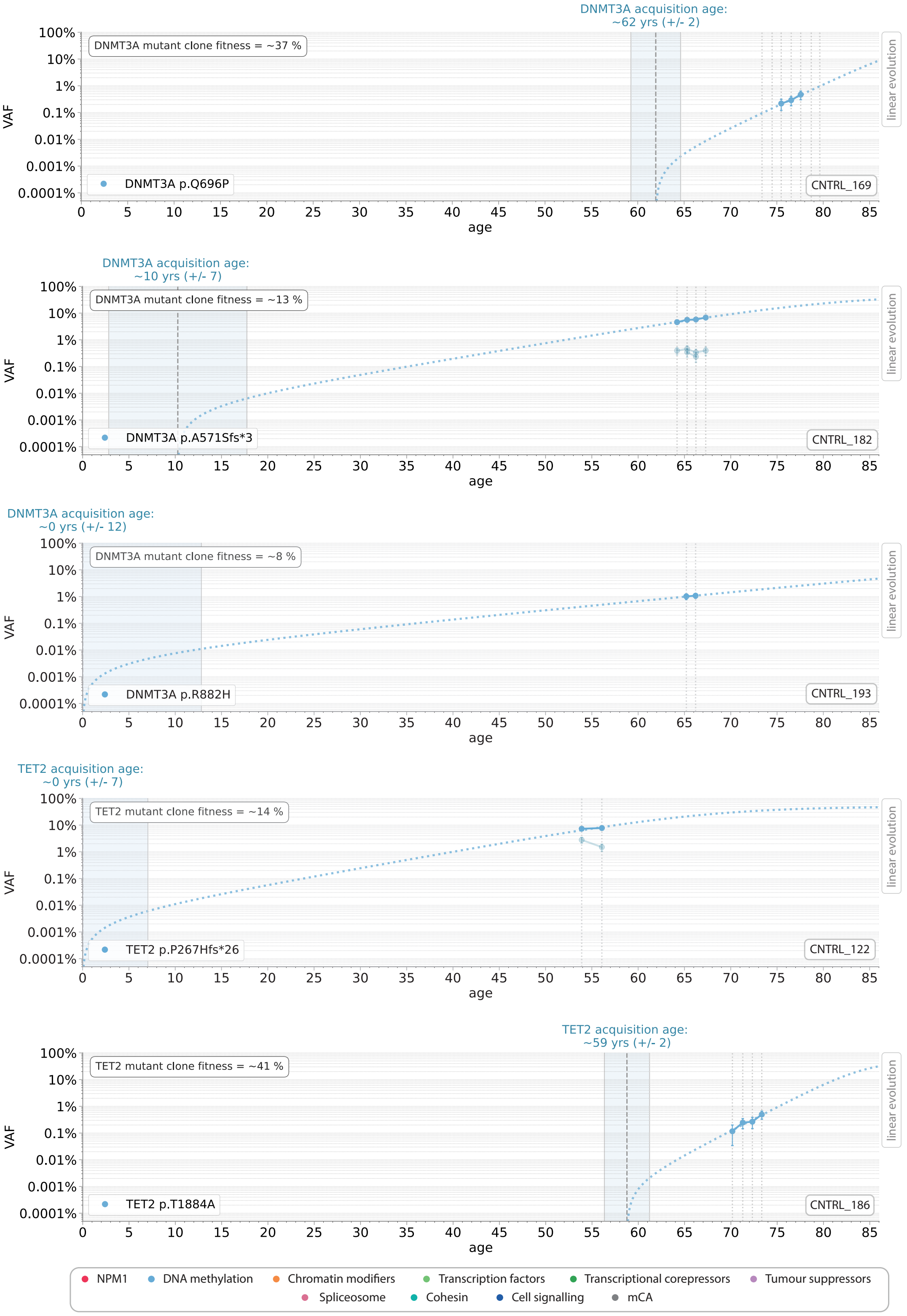


Figure D.14 Estimation of acquisition age and fitness for mutations showing a linear evolution pattern: Part 5. Grey vertical lines indicate blood sample timepoints. Trajectories are coloured according to their class gene class (see legend). Dash coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches. The error measurement shown for the acquisition ages is $\pm 1/s$. Other mutations present in the sample are shown as faded datapoints.

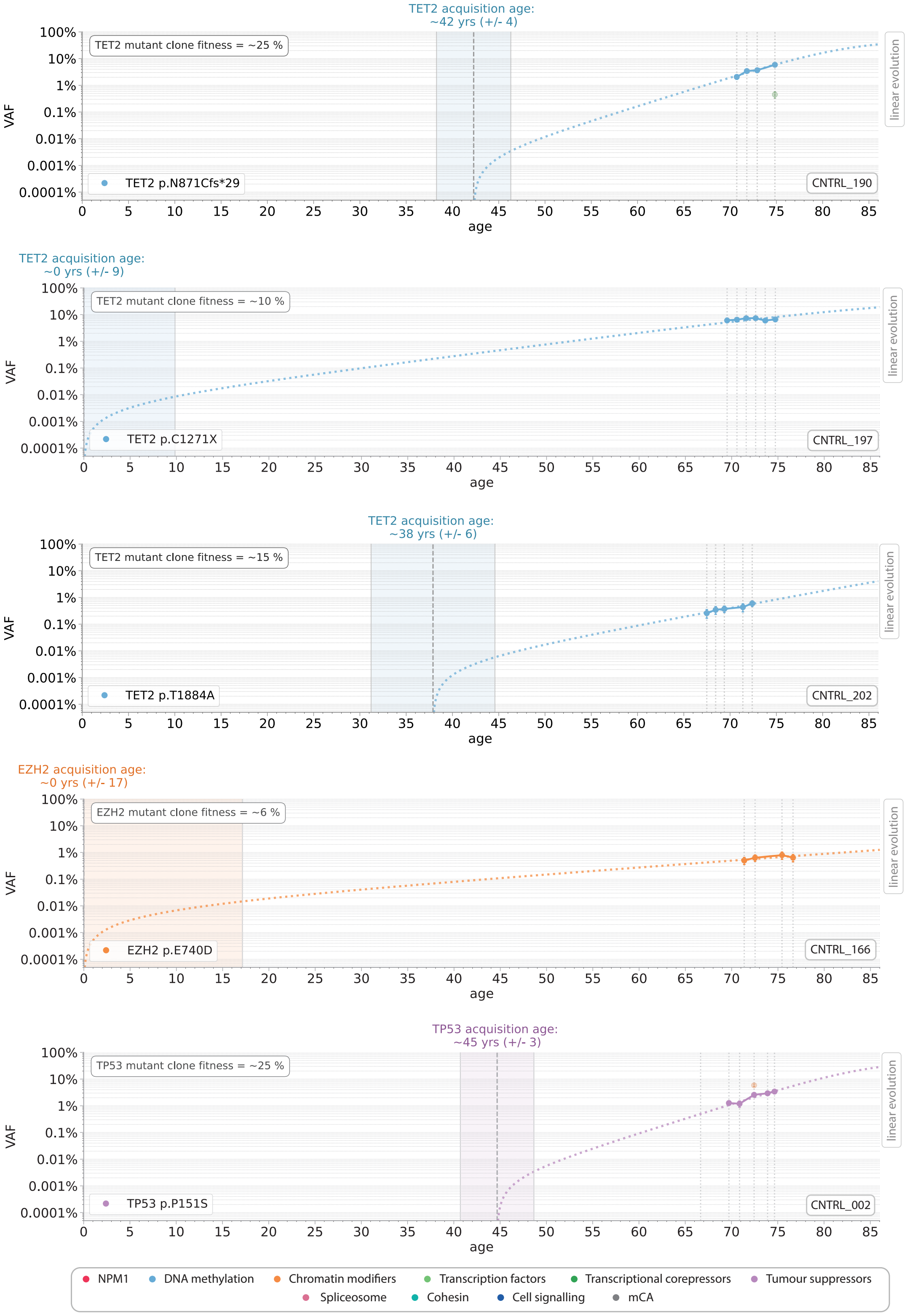


Figure D.15 Estimation of acquisition age and fitness for mutations showing a linear evolution pattern: Part 6. Grey vertical lines indicate blood sample timepoints. Trajectories are coloured according to their class gene class (see legend). Dash coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches. The error measurement shown for the acquisition ages is $\pm 1/s$. Other mutations present in the sample are shown as faded datapoints.

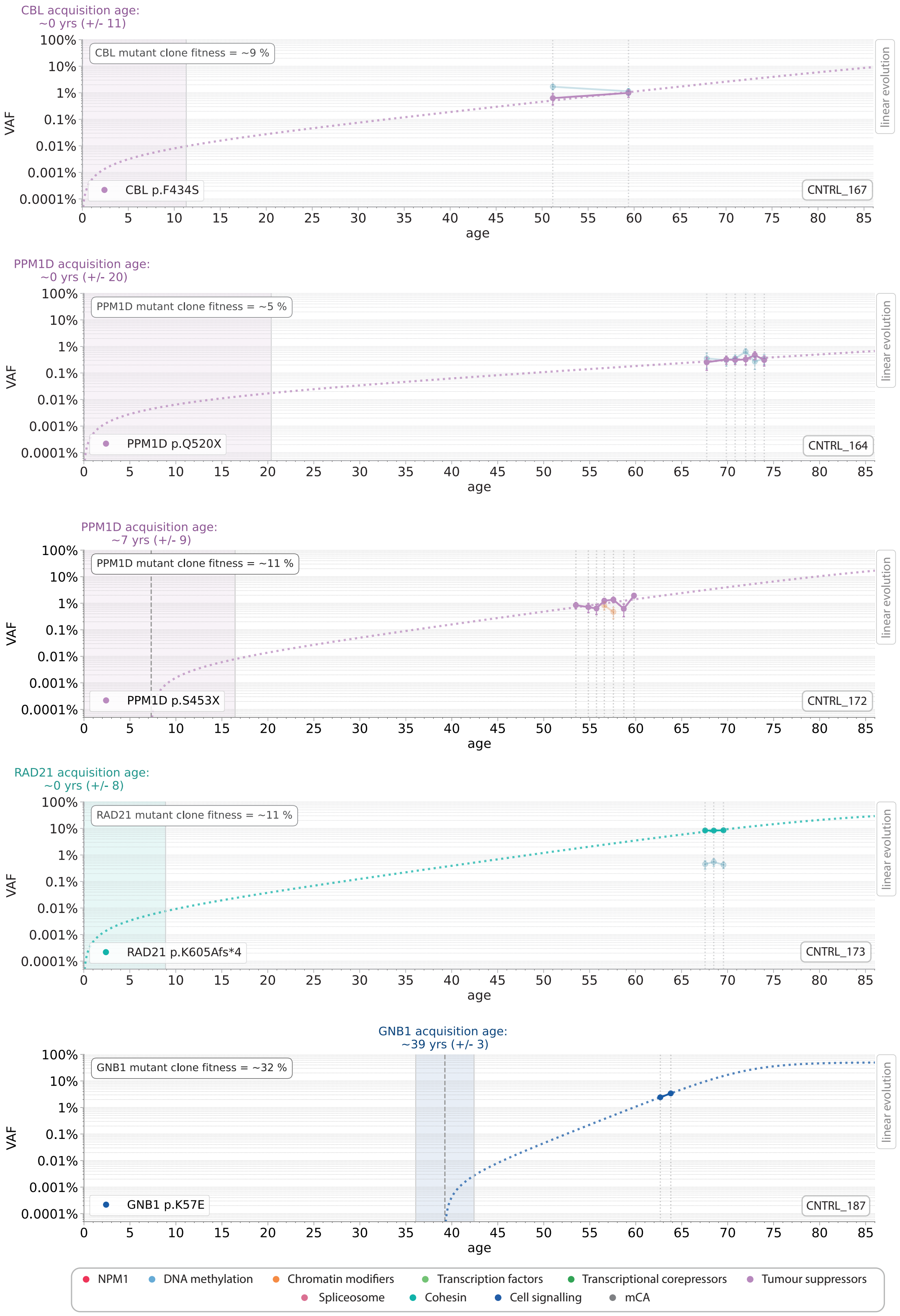


Figure D.16 Estimation of acquisition age and fitness for mutations showing a linear evolution pattern: Part 7. Grey vertical lines indicate blood sample timepoints. Trajectories are coloured according to their class gene class (see legend). Dash coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches. The error measurement shown for the acquisition ages is $\pm 1/s$. Other mutations present in the sample are shown as faded datapoints.

D.3 Mutation acquisition age and fitness for late evolution samples

Supplementary material for Section 5.3.2

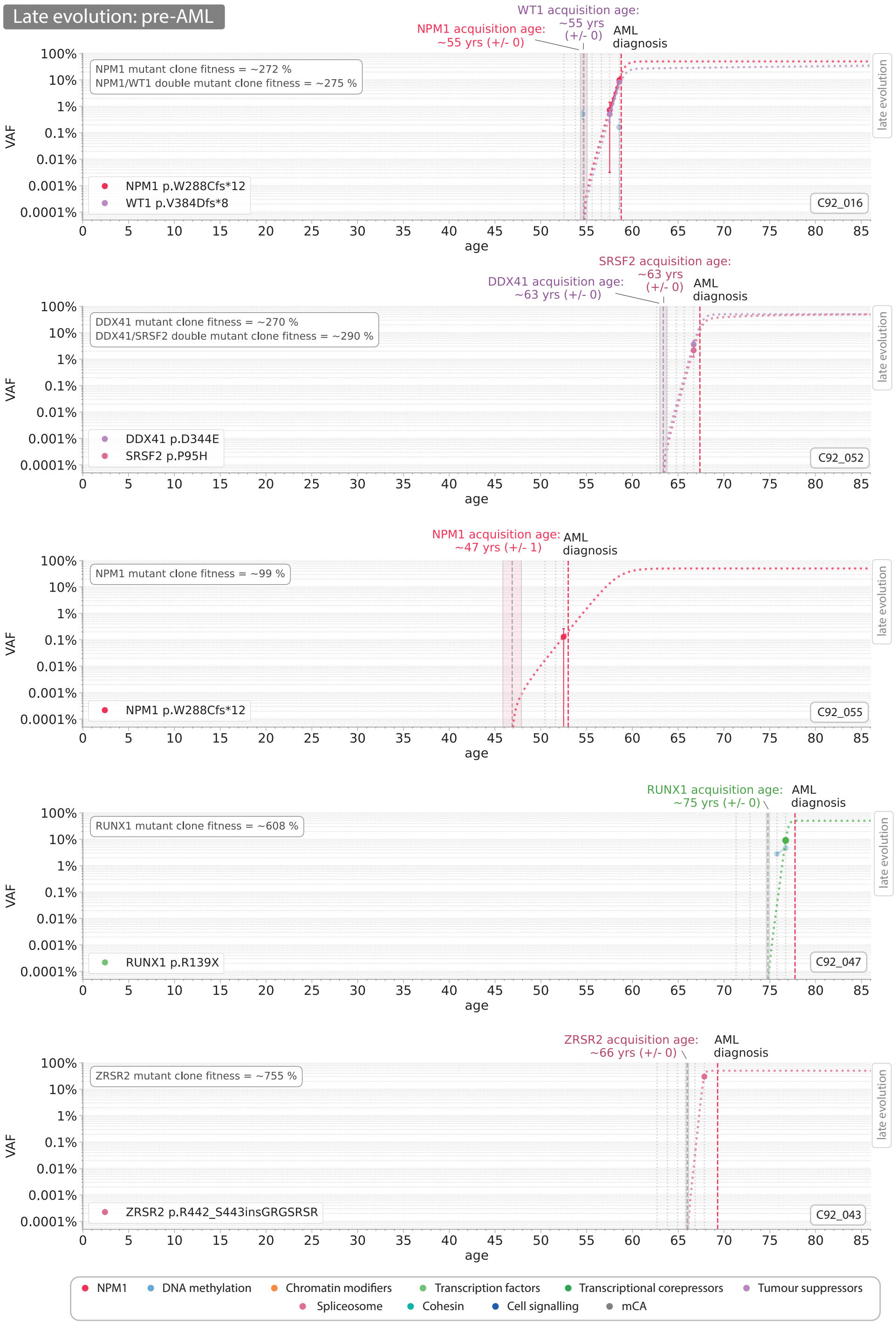


Figure D.17 Estimation of acquisition age and fitness for mutations showing a late evolution pattern: Part 1. Grey vertical lines indicate blood sample timepoints. Trajectories are coloured according to their class gene class (see legend). Dash coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches. The error measurement shown for the acquisition ages is $\pm 1/s$. Other mutations present in the sample are shown as faded datapoints.

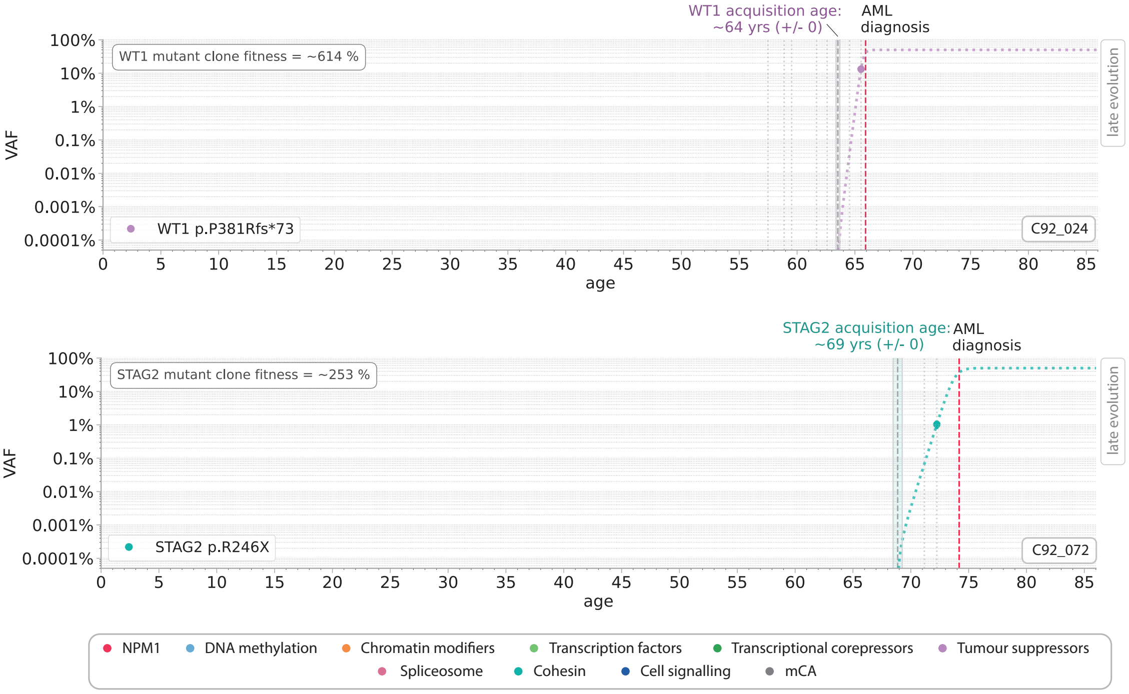


Figure D.18 Estimation of acquisition age and fitness for mutations showing a late evolution pattern: Part 2. Grey vertical lines indicate blood sample timepoints. Trajectories are coloured according to their class gene class (see legend). Dash coloured lines indicate the extrapolated trajectories inferred using maximum likelihood approaches. The error measurement shown for the acquisition ages is $\pm 1/s$. Other mutations present in the sample are shown as faded datapoints.