PATTERNS OF SOMATIC MUTATIONS IN NORMAL CELLS

Freek Martijn Manders

ISBN: 978-90-393-7537-2
DOI: https://doi.org/10.33540/1622
Copyright: Freek M. Manders, all rights reserved
Cover illustration and layout: Freek M. Manders, all rights reserved
Printing: Gildeprint - Enschede

FSC www.fsc.org MIX Paper from responsible sources

Patterns of somatic mutations in normal cells

Patronen van somatische mutaties in normale cellen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

21 februari 2023 des middags te 4.15 uur

door

Freek Martijn Manders

geboren op 3 januari 1995
te Veghel

Promotoren:
Prof. dr. E.P.J.G. Cuppen
Prof. dr. F.C.P. Holstege

Copromotor:
Dr. R. van Boxtel

Beoordelingscommissie:
Prof. dr. G.J.P.L. Kops
Prof. dr. J.A.M. Borghans (voorzitter)
Prof. dr. B. Snel
Prof. dr. L.F.A. Wessels
Dr. P.P.C.W. Kemmeren

TABLE OF CONTENTS

Chapter 1 Introduction: The dynamics of somatic mutagenesis during life in humans **6**

Chapter 2 Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis **28**

Chapter 3 MutationalPatterns: the one stop shop for the analysis of mutational processes **66**

Chapter 4 Mutation accumulation in mitochondrial DNA of cancers resembles mutagenesis in normal stem cells **106**

Chapter 5 Investigation of single-cell genomes at nucleotide resolution using the PTA Analysis Toolkit (PTATO) **138**

Chapter 6 General discussion **186**

Addendum Nederlandse samenvatting **210**
Author contributions per chapter **214**
List of publications **215**
Dankwoord **217**
Curriculum Vitae **220**

Chapter 1

Introduction: The dynamics of somatic mutagenesis during life in humans

Freek Manders[1], Ruben van Boxtel,[1] and Sjors Middelkamp[1,*]

[1]Princess Máxima Center for Pediatric Oncology and Oncode Institute, Heidelberglaan 25,
3584CS Utrecht, The Netherlands
[*]Corresponding author: s.h.a.middelkamp-5@prinsesmaximacentrum.nl

Modified from: Frontiers in Aging, 2021

Abstract

From conception to death, human cells accumulate somatic mutations in their genomes. These mutations can contribute to the development of cancer and non-malignant diseases and have also been associated with aging. Rapid technological developments in sequencing approaches in the last few years and their application to normal tissues have greatly advanced our knowledge about the accumulation of these mutations during healthy aging. Whole genome sequencing studies have revealed that there are significant differences in mutation burden and patterns across tissues, but also that the mutation rates within tissues are surprisingly constant during adult life. In contrast, recent lineage-tracing studies based on whole-genome sequencing have shown that the rate of mutation accumulation is strongly increased early in life before birth. These early mutations, which can be shared by many cells in the body, may have a large impact on development and the origin of somatic diseases. For example, cancer driver mutations can arise early in life, decades before the detection of the malignancy. Here, we review the recent insights in mutation accumulation and mutagenic processes in normal tissues. We compare mutagenesis early and later in life and discuss how mutation rates and patterns evolve during aging. Additionally, we outline the potential impact of these mutations on development, aging and disease, which leads to a description of the aims and scope of this thesis

Introduction

Virtually every cell in the body contains a unique set of changes to the genome due to the accumulation of somatic mutations during life. Some of the mutations cells acquire during life can contribute to the development of age-associated diseases, such as cancer[1]. Mutations can result from errors made during DNA replication or from unrepaired or incorrectly repaired DNA damage. Each mutational process leaves characteristic patterns of mutations, or “mutational signatures”, in the genome, which can be identified by systematically studying mutation spectra[2,3].

Somatic mutations have historically been hard to detect, because they are often present in only a tiny fraction of an individual's cells resulting in a low variant allele frequency[4]. As a result, most somatic variants are not detected by regular bulk tissue sequencing technologies. Notable exceptions to this are somatic variants in cancer. Since cancers grow out from a single cell, all the somatic variants in that original cell will be clonally present in the cancer. Somatic mutations in cancer have been extensively studied[5]. However, to better understand which somatic mutations and mutational processes contribute to cancer, somatic mutations in cancer need to be compared to somatic mutations in normal, pre-cancer tissues for both the nuclear and mitochondrial genome.

In the last few years, technological developments in DNA sequencing methods and bioinformatic approaches have enabled the detailed study of somatic mutations in normal tissues. *In vitro* expansion of a single stem cell, followed by whole genome sequencing of the clone, enables highly accurate characterization of the genome of a single cell[6]. Additionally, somatic variants can be identified by (deep) sequencing of natural occurring clonal patches in healthy tissues using low-input sequencing[7]. A disadvantage of these methods is that they are limited to cells with self-renewal capacity. Direct single-cell sequencing after whole genome amplification and single-molecule duplex sequencing are methods that can also be applied to non-dividing cells. Until recently these methods had a relatively low accuracy in mutation detection, but new studies using novel technical and bioinformatic innovations claim to have significantly reduced their error rates[8–10]. While novel whole genome amplification techniques have reduced error-rates, they still contain a sizeable number of artifacts, necessitating the use of stringent bioinformatics analyses to achieve high-quality data. For one of these techniques, primary template-directed amplification (PTA), we developed a comprehensive bioinformatics analysis pipeline, which we investigated in more detail in the work described in **chapter 5**.

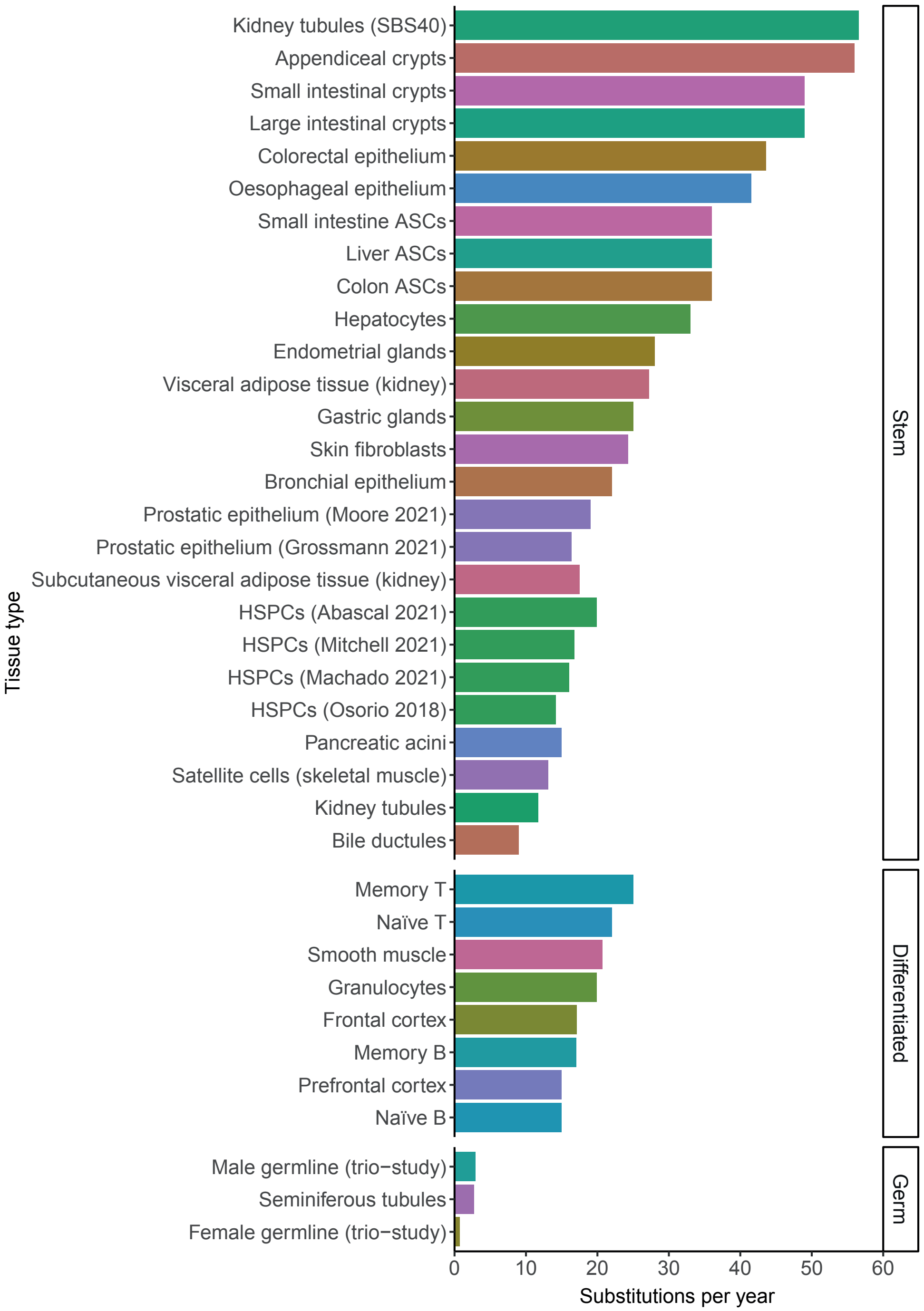


Fig. 1: The number of substitutions per year for different tissue types.
When the mutation rate of a tissue has been determined multiple times, they are distinguished by the last name of the first author and the publication year. The color indicates the tissue type. The mutation rates in this figure may be influenced by technical differences between the studies, which may explain some of the small differences between tissues. HSPCs; Hematopoietic stem and progenitor cells. ASCs; adult stem cells. Source of mutation rates: Kidney tubules, Subcutaneous visceral adipose tissue (kidney) and Visceral adipose tissue (kidney)[25]. Appendiceal crypts, small intestinal crypts, large intestinal crypts, gastric glands, pancreatic acini, bile ductules and seminiferous tubules[20]. Colorectal epithelium[26]. Esophageal epithelium[27]. Small intestine ASCs, liver ASCs and colon ASCs[15]. Hepatocytes[24]. Endometrial glands[29]. Skin fibroblasts[21]. Bronchial epithelium[17]. Prostatic epithelium[18,20]. HSPCs[8,19,23,55]. Satellite cells[16]. Memory T, Naïve T, Memory B, Naïve B[55]. Smooth muscle, granulocytes, Frontal cortex[8]. Prefrontal cortex[9]. Male and female germlines[31].

Many studies have recently applied single cell DNA sequencing to characterize somatic mutation accumulation in various tissues of healthy human donors across a wide range of ages. Here, we provide an overview of the findings of these papers and discuss how somatic mutagenesis evolves during life. First, we review how mutations accumulate linearly with age. Next, we review how activity of specific mutagenic processes can be analyzed using mutation data and how these contribute to the differences in mutation accumulation observed between different adult tissues. Building on this knowledge of mutagenesis in adults, we show how mutagenesis is divergent early in life before birth. Additionally, we discuss the impact of somatic mutations and how this is different between mutations that occur early and later in life. Finally, we discuss the challenges with identifying mutations in mitochondrial DNA, which are ignored in many studies, even though they have been associated with diseases like aging and cancer[11–14].

Adult tissues accumulate mutations linearly with age

A mutation that is acquired in one cell, will be propagated to all of its progeny. It has become clear in recent years that the somatic mutation burden increases remarkably linearly with age in single cells in normal tissues[15]. So far, this linear mutation accumulation is confirmed in stem cells of all studied normal tissues including liver, small intestine, large intestine, lung, skin, blood, esophagus, muscle, kidney, adipose tissue, endometrium, bile duct, stomach, prostate, pancreas, appendix and bladder[15–29]. The number of mutations that accumulate in different tissues ranged from 9 substitutions per year in bile ductular cells to 56 substitutions per year in appendiceal crypts[20]. Mutation rates in other tissues fell within this range, showing that while there are differences between tissues, they all fall within a single order of magnitude (Figure 1). Female and male germ cells acquire only 0.74 and 2.7 mutations per year, showing that the mutation rate in somatic cells is much higher than in germline cells[20,30,31]. The average mutation rates within tissues appear to be relatively constant during adult life. However, individual cells may have mutation burdens divergent from the average burden in the tissue due to different exposures to endogenous or exogenous mutagenic processes and due to differentiation, as will be discussed in the next sections.

Analyzing mutational processes

Mutational signature analysis can be used to infer past activity of or exposure to mutagenic processes[3,32]. Mutational signatures are generally created using a computational data reduction method called non-negative matrix factorization (NMF)[33,34]. This approach assumes that the mutational landscape of a single sample has been shaped by the activity of multiple mutational processes, making it impossible to disentangle the contribution of each individual processes when assessing a single sample. However, the contribution of these processes will differ between samples, as mutagenic exposure varies across individuals or even single cells, which allows for identifying recurrent patterns when assessing large numbers of samples. These patterns are based on the contribution of different mutation types to a sample. For example, for single base substitution signatures, mutations are divided into the 6 types of base substitutions and their direct 5' and 3' flanking bases resulting in 96 trinucleotide changes in which the middle base is mutated[35,36]. We identify only 6 types of base substitutions, because after DNA replication a G>A on the "-" strand will result in a C>T on the "+" strand, making these two mutation types generally undistinguishable, although some mutational processes do result in an observable strand asymmetry. Next, a count matrix M is created containing the number of mutations per sample per mutation type. This matrix is then factorized into a signature matrix S and an exposure matrix E, so that $M \approx S * E$[34]. This is somewhat analogous to PCA, where a dataset is factorized into a matrix containing the location of samples within the principal components and another matrix containing the directions of the principal components. Next to identifying novel mutational signatures, it is also possible to determine the contribution of signatures to single samples via a process known as "signature refitting" or to determine the similarity of samples to predetermined signatures by calculating the "cosine similarity"[36,37].

The number of known mutational signatures has grown rapidly over the last few years. The earliest work on mutational signatures identified only 5 signatures on 21 breast cancers[35]. The following year this was extended to 22 signatures and the latest version of the catalogue of somatic mutations in cancer (COSMIC) (v3.2) contains 78 signatures[2,36]. While signatures were initially only defined for single base substitutions, COSMIC now also contains signatures for double base substitutions and indels[36]. Attempts to identify mutational signatures for other mutation types like larger structural variants and copy numbers have also been made[38–40]. Next to identifying signatures with NMF in cancer samples, some signatures have also been directly identified or confirmed via *in vitro* treatments[41–43]. Many mutational processes can be active in multiple tissues; however, they can be somewhat different between these tissues, possibly because of slight differences in the DNA-repair and the turnover rate of the tissue. Similarly, mutational processes also differ between species[44]. Mutagenic processes can be investigated using more methods than only mutational signatures. Different types of mutations and their context can for example be compared directly between samples. Additionally, kataegis, which is a process of localized hypermutation caused by APOBEC, can be detected by identifying small regions with a high mutation density[35,45]. Another example is the strand asymmetry that is caused by lesion segregation[46]. This occurs when a mutational process causes many mutations, for example C>T, during a single cell cycle. After DNA replication the C>T mutations will be present on the + strand of one copy and the – strand of the other copy. After mitosis, a daughter cell will inherit the copies of both alleles in a 1:2:1 ratio, which means that there is a 50% change that the mutations in each part of the genome are all on the same strand. Lesion segregation can be identified when a cell or descendent of a cell in which it has occurred is sequenced and the strand of the mutations is visualized.

To perform the types of analyses described above, we developed the second version of the MutationalPatterns package, which we describe in **chapter 3**.

Tissue-specific mutational processes in adult stem cells

Mutational signature analysis has indicated that some mutagenic processes are active in all tissues, whereas some are tissue- or exposure-specific. SBS1 and SBS5, which reflect life-long activity of "clock-like" mutational processes, which cause mutations at a steady rate, were found in all cell types. SBS1 mutations are thought to be caused by spontaneous deamination of methylated cytosine residues into thymine. In contrast, the cause for SBS5 mutations is unknown, but likely represents a collective of endogenous background mutational processes[2,35]. REV1 has also been associated with this signature[47]. While most, if not all, tissues gradually accumulate SBS1 and SBS5 mutations throughout life, their ratios differ between tissues. Differences in cell turnover rate between tissues have been suggested to be one possible cause for this[15,20,48,49].

Some cells also showed contributions of additional mutational signatures caused by both exogenous and endogenous factors, which explain part of the variation in the mutation rate and spectra between tissues. Most skin fibroblasts and melanocytes, for example, show contribution of SBS7, a signature caused by UV-radiation[21,50]. Similarly, kidney tubule cells with contribution of SBS40, in this case possibly caused by formaldehyde and alkylating agents, were found to accumulate 56.6 SNVs per year whereas cells lacking this signature only accumulated 11.7 SNVs per year[25].

SBS16, which is associated with alcohol consumption, could be found in cells of the esophagus[20,27]. Colibactin produced by a specific common *E. coli* strain was found to cause SBS88 mutations in some colon crypts[20,26,42]. SBS2 and SBS13, which are associated with activity of endogenous APOBEC cytosine deaminases, have been found in multiple cell types including lung, colorectal and small intestinal cells[17,20,26]. These signatures are caused by sporadic bursts of mutagenesis in a subset of cells in a tissue[51]. Overall, it is clear that SBS1 and SBS5 are present in almost all cell types and that additional tissue-specific mutational processes can result in an increased mutational burden.

Mutagenesis in post-mitotic and differentiated cells

It is likely that there are differences in mutagenesis between stem cells and non-dividing, differentiated cells. Stem cells could be expected to be protected from mutagenesis, because they are long-lived and can self-renew in order to regenerate tissues throughout life, which is not the case for most post-mitotic or fully differentiated cells[52]. On the other hand, post-mitotic cells will not accumulate errors made during DNA replication. So far, somatic mutations have been mostly characterized in proliferating stem and progenitor cells due to technical limitations. In the last few years, technical innovations have also enabled a more accurate detection of somatic variants in non-dividing cells. One study using single-cell whole genome amplification found a higher mutation rate in differentiated hepatocytes compared to liver stem cells[53]. Differentiated granulocytes were found to have a slightly, not significantly, increased mutational load compared to hematopoietic stem cells[8]. In contrast, both naïve and memory T-lymphocytes showed an increased mutation rate of 22 and 25 SNVs per year compared to the 16 SNVs per year that this study found in hematopoietic stem cells. Additionally, both memory B- and T-lymphocytes showed an increased mutation load irrespective of age, likely caused by somatic hypermutation, while naïve lymphocytes did not[54,55]. Interestingly, several studies found that post-mitotic brain neurons accumulate mutations with age at a similar rate (14.7-17.1 per year) as stem cells of other tissues, even though they do not replicate[8,9,56].

Most somatic mutations in short living differentiated cells are likely acquired in the stem cell ancestors of these cells. Their relatively short lifespan might prevent differentiated cells from building up a strongly elevated mutation burden after differentiation, even if their mutation rate would be much higher. In addition to an increased mutation rate, it is also possible that the process of differentiation itself could lead to a single burst of mutation accumulation. Estimating the precise mutation rate in *in vivo* differentiated cells is therefore challenging. Overall, the first single-cell and single-molecule sequencing studies on differentiated and post-mitotic cells, suggest that the mutation burden in these cells is either not or only modestly increased as compared to stem cells in the same tissues.

Somatic mutation rates are strongly elevated in prenatal cells

There are two main technical issues with detecting somatic mutations in prenatal cells. First is the low total mutation load of fetal cells. Because false positives in whole genome sequencing are generally not dependent on the number of true positives, the false positive rate can increase[57,58]. These factors make it necessary to filter very stringently for true mutations and make it more difficult to correctly identify mutational processes. Another issue in detecting somatic variants in prenatal cells, is that any variant that occurred pre-gastrulation can be (sub)-clonally present in bulk tissues[59,60]. The common method of using a bulk control when performing whole genome sequencing will thus not detect these early somatic mutations. This issue can be solved by comparing the genomes of multiple single normal cells from an individual and reconstruct a phylogenetic tree by assessing mutations that are shared in some, but not all cells. With this approach each cell functions as a control for the other cells.

We and others have used this approach to study prenatal mutation accumulation in stem cells[61,62]. Interestingly, these analyses showed that the somatic mutation rate during fetal development is strongly increased compared to the post-natal rates. For example, in hematopoietic stem cells of newborns an average of 40 somatic SNVs were found nine months after conception, while only 14-17 SNVs are gained in adult stem cells per year[19,23]. This increased mutation rate before birth was confirmed by genomic analyses of various fetal tissues[61–64]. The mutation rate in fetal liver, intestine and hematopoietic stem cells was shown to be fivefold higher as compared to the rate in adult tissues. The mutation rate in fetal mitotic neuronal progenitor cells was even suggested to be 50-fold higher than in other postnatal tissues[63], although more recent studies could not confirm this high mutation rate[8,9].

Additional evidence for an elevated prenatal mutation rate was provided by retrospective lineage tracing studies of cell lineages in various tissues. Since a mutation that arose in one cell will be inherited by all its progeny, somatic mutations can be used as genetic barcodes, which enables the reconstruction of phylogenies and the identification of early embryonic cell divisions (Figure 2)[65]. By assessing mutations that are shared between different cells of the same individual we and others have shown that especially in the first two to three cell divisions (up to the 8-cell stage) the mutation rate is relatively high with roughly 2 to 3 mutations per cell division[21,62,66]. In addition to the high mutation rate in the first embryonic divisions, it was previously reported that chromosomal instability, characterized by high levels of chromosomal missegregation and abnormalities leading to mosaicism, is common in early (cultured) human embryos[67].

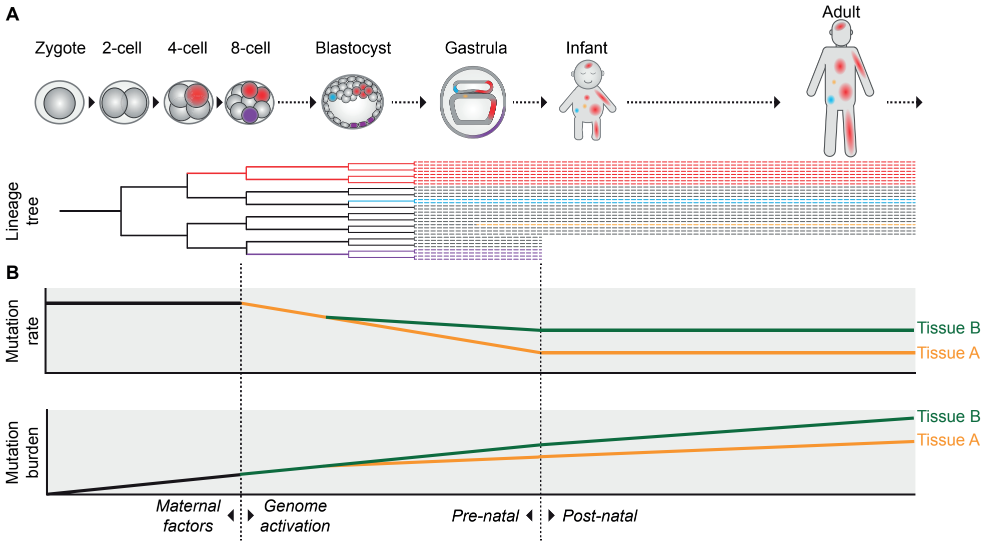


Figure 2: Proposed model of the dynamics of somatic mutagenesis during life.
Schematic overview depicting the distributions and rates of somatic mutations. **a** Mutations arising early in development can be propagated to many cells of multiple tissues, as indicated by the red cell lineage. Due to this wide distribution, mutations arising early in life can have a strong potential impact on development and disease. Mutations acquired later in life are usually only inherited by a small number of cells (the blue- and orange-colored cells). Some early mutations, depicted by the purple lineage here, may also end up in extra-embryonic cell lineages not contributing to the embryo proper. **b**The somatic mutation rate is especially high in the first embryonic cell divisions. After genome activation, the mutation rate decreases. It is unclear if this decrease is gradual (as depicted) or more abrupt, but the mutation rate probably remains relatively high compared to the postnatal mutation rate. After birth, the somatic mutation rate appears to stay remarkably constant during aging, leading to a gradual linear mutation accumulation. Variance in the mutation rate between tissues leads to a tissue-specific mutation burden. In some tissues (such as intestine), the tissue-specific mutation patterns already arise early in embryogenesis, whereas in others (such as liver) these patterns start to emerge only after birth.

Several non-exclusive factors contributing to the high mutation rates in early human embryos have been proposed. The first cell cycles after fertilization are relatively fast, leaving little time for proper DNA repair[68]. Transcription is not yet active in the first cell divisions. This precludes high-fidelity transcription-coupled DNA repair and makes DNA repair entirely dependent on maternally inherited factors, which are diluted with every cell division[69]. DNA damage can be caused by chromatin remodeling after fertilization or may be inherited from the sperm cell[70]. Cell cycle and DNA damage checkpoints are relaxed and apoptosis is prevented, raising tolerance to DNA damage and mutagenesis[68,71,72]. After the 8-cell stage, coinciding with activation of transcription, the mutation rate drops to less than 1 mutation on average per cell division[21,62]. The elevated mutation rate, leading to the presence of dozens of somatic mutations in each cell at birth, may be the cost of the rapid growth required during embryonic development (Figure 2B). The pre-natal mutation rate is increased even more in fetuses with Down-syndrome, which we investigated in more detail in the work described in **chapter 2**.

Divergence of tissue-specific mutation patterns early in life

The mutation patterns and rates that are specific for each tissue in adults must emerge at a certain, currently unclear, moment during development. As both cellular functions and exposure to exogenous agents are different before and after birth, it can be expected that mutation accumulation also differs between those phases in life. The earliest embryonic mutations show clock-like mutational signatures that are also common in most adult tissues, namely SBS1 and SBS5[21,62,66]. Later during development, it has been shown that fetal intestinal cells show the same specific mutation patterns as in adult intestinal cells already at 13 weeks after fertilization[64]. In contrast, liver stem cells and hematopoietic stem cells show different mutation patterns before and after birth[61,64]. Fetal liver stem cells show a high contribution of SBS18, which is linked to oxidative stress-induced mutagenesis and interestingly was also found at high levels in fetal neural progenitors[43,63,73]. These findings show that mutation patterns start to diverge between tissues already early in development, but the precise moment appears to be tissue specific.

The impact of somatic mutations on disease and aging

Somatic mutations can influence disease and aging in multiple ways. The most well-known impact of somatic mutations is their involvement in cancer[74]. Somatic mutations can also lead to non-cancerous, but potentially harmful clonal outgrowth, with clones sometimes replacing entire tissues, for example in clonal hematopoiesis[75–77]. As has been recently reviewed in detail, accumulation of somatic mutations might also impact aging for example by affecting the functioning of cells by influencing tightly controlled gene-regulatory networks and increasing cell-to-cell transcriptional heterogeneity (transcriptional noise)[78]. This is further supported by the observation that the mutation load at the end of life is similar between different mammals with wildly different lifespans[79]. Defects in DNA repair have also been associated with accelerated aging[80]. However, it has recently been shown that a massively increased somatic mutation rate, in this case due to germline *POLE/POLD1* mutations, does not necessarily lead to accelerated aging[81]. Further studies comparing young and old tissues are required to elucidate the complicated effects of somatic mutations on normal and aberrant aging.

Early mutations: Small numbers, but large effect

In adults, only a relatively small fraction of all mutations in each cell was acquired prenatally during early development. While these early mutations are less numerous, they are shared by a high fraction of the individual's cells (Figure 2A)[59,66]. As a result, some of these mutations might be clinically relevant, as somatic mosaicism can underlie genetic diseases and cancer[82–85]. For example, somatic mosaicism has been associated with autism spectrum disorders as well as other neurological disorders[86–88]. As many of these disorders originate early in life, especially mutations arising in the earliest phases of development may have a pathogenic impact[89].

A cancer driver mutation arising early in development can be propagated to a large fraction of cells. As a result, a relatively large population of cells will be vulnerable to developing into a malignancy via further hits. Consistent with this, driver mutations have been found that originated decades before the development of cancer, some of which likely emerged during fetal development or early childhood[90,91]. In addition to adult cancers, pediatric cancers are likely often caused by somatic mutations occurring during development[92]. It has been shown that 1% of newborns already have detectable driver genomic rearrangements in some of their blood cells, but these only lead to cancer in a very small minority of cases[93].

Early mutations may also impact spontaneous abortions. Less than half of all human conceptions are thought to lead to a live birth. This is at least partly due to somatic aneuploidies and copy number changes occurring in the first cell divisions, but it is not unlikely that in some cases SNVs also play a role[94].

Mutations that occur early in development can end up in the germline and thereby propagate as de novo germline variants in a person's offspring. Early mutations occurring during development can likely explain a sizable fraction of de novo germline variants, because of the low mutation rate of germline cells[95]. These de novo germline variants can cause (neuro-)developmental disorders and other diseases[96–98].

Mitochondria

Compared to samples from children, the mutation load in mitochondria is even lower. Mitochondria contain their own 16.6kb-long DNA of which more than ten copies can be present per mitochondrion[99,100]. More than a thousand of these mitochondria can be present in a single cell[11,101]. Unequal replication of mitochondrial genomes and a random division of mitochondrial genomes after mitosis can change the variant allele frequency of mutations, a phenomenon known as heteroplasmy[102]. Because of these properties, identifying mitochondrial mutations can be challenging. Other issues with identifying mitochondrial mutations are the circular nature of the mitochondrial genome and nuclear mitochondrial sequences, which are insertions of the mitochondrial genome into the nuclear genome. Because of these issues, the mitochondrial genome is often ignored when performing whole genome sequencing. However, mutations in the mitochondrial DNA have been suspected to be related to the development of a variety of neurodegenerative diseases, aging and the onset or progression of cancer[11–14]. In the work described in **chapter 4** we show that mitochondrial genomes of normal cells accumulate mutations with age just like the nuclear genome and we show that most mitochondrial mutations in cancer are the result of premalignant normal mutagenesis.

Discussion

Recent studies have revealed that within a healthy tissue the accumulation of somatic mutations in stem cells occurs at a remarkably constant rate during life. Despite significant differences in function, cell turnover, exposures to mutagens and cancer incidence, but also technical differences between studies, the variation in mutation rates and patterns between stem cells of different tissues is surprisingly modest. Most somatic mutations in stem cells of normal tissues are characterized by different contributions of mutational signatures SBS1 and SBS5, but some cells show contributions of signatures caused by tissue-specific endogenous and exogenous exposures. The mutation burden and in many cases also the clonality of tissues increases with aging, but there seem to be no apparent differences in mutation rates and patterns between stem cells of old and young individuals. The elevated mutation rate before birth forms a notable exception (Figure 2). It appears that the rapid growth during fetal development comes at the cost of decreased genomic stability.

Differentiation, which can change the cell's functions, self-renewal capacity and potentially also the exposure to mutagens, could be expected to lead to changes in mutation rates and spectra. So far, the first genomic studies of differentiated cells suggest that the effect of differentiation on mutagenesis is only modest, possibly due to the relatively short lifespan of differentiated cells. More single-cell studies of both *in vivo* and *in vitro* differentiated cells of various tissues are required to elucidate the precise role of differentiation on mutation accumulation. Intriguingly, post-mitotic neurons also accumulate mutations at a linear rate similar to stem cells of other tissues. This shows that cell division and DNA replication do not necessarily have to be the main drivers of somatic mutagenesis in all cells. The linear mutation accumulation suggests that mutagenesis in these non-dividing cells, likely caused by endogenous DNA damage followed by erroneous DNA repair, also occurs at a relatively steady rate.

Somatic mutations can result in cancer and play a role in other diseases. They have also been associated with aging, though more research is needed to strengthen these claims. It seems likely that mutations arising early in life, even though they are less numerous than somatic mutations occurring during adulthood, regularly impact disease. The rapidly increasing amount of genomics data will help to further elucidate the relative impact of these early-life mutations compared to the ones arising later in life. Similarly, the impact of mitochondrial mutations will also become clearer.

Thesis scope and outline

In the work described in this thesis, we aimed to improve, develop, and apply new methods to investigate mutational processes to better understand the initiation and development of cancer, with an emphasis on pediatric cancers.

In this general introduction (**chapter 1**), we focused on mutation accumulation in normal cells.

In the work described in **chapter 2**, we characterized the somatic mutation accumulation in human fetal hematopoietic stem and progenitor cells (HSPCs) and intestinal stem cells. We found an increased mutation rate in fetal HSPCs compared to post-natal cells. The mutation load was even higher in fetal trisomy 21 (T21) cells, which may contribute to the increased risk for leukemia in children with Down syndrome. Mutational signature analysis showed that the same mutational processes explain mutation accumulation in karyotypically normal and T21 cells as well as in Down syndrome-associated leukemia samples.

Chapter 3 describes the development of a new version of the MutationalPatterns R package, making use of methods developed in the previous chapter, which can be used to investigate mutational processes and applied it on several datasets. The new version of the package supports more mutation types and can be used to perform stricter signature refitting, regional pattern analyses, and lesion segregation analyses. We applied the package on cell-lines with DNA-repair gene knockouts and identified the mutational signatures contributing to them.

The investigation described in **chapter 4** goes beyond the nuclear genome by investigating mutations in the mitochondrial genomes of normal single cells. We found that mitochondrial mutations accumulate with age and that most of the mutations found in cancer are the result of healthy regular mutation accumulation in healthy cells. Finally, we showed that chemotherapy treatment does not impact the mitochondrial mutation load or mitochondrial DNA copy numbers of most cells.

In the work described in **chapter 5**, we developed a comprehensive analysis pipeline for primary template-directed amplification (PTA), allowing us to investigate non-dividing cells, instead of only being restricted to stem cells. The pipeline can create quality control plots, filter SNVs and indels from artifacts and filter and integrate structural and copy number variants. We applied the pipeline on both healthy blood cells and AML cells of a pediatric AML patient to time mutational processes and the acquisition of driver mutations.

In **chapter 6** our findings are summarized and discussed in a broader perspective. Recommendations for future research directions are also presented.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

F.M., R.v.B. and S.M. jointly wrote the manuscript. R.v.B. and S.M. supervised the project. All authors approved the manuscript for publication.

Funding

This study was financially supported by a VIDI grant from the Netherlands Organisation for Scientific Research (NWO no. 016.Vidi.171.023) to R.v.B.

References

1. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science (80-). 2015;349(6255):1483–9.
2. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug;500(7463):415–21.
3. Koh G, Degasperi A, Zou X, Momen S, Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer. 2021;21(10):619–37.
4. Dou Y, Gold HD, Luquette LJ, Park PJ. Detecting Somatic Mutations in Normal Cells. Trends Genet. 2018;34(7):545–57.
5. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93.
6. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC, et al. The origin and evolution of mutations in acute myeloid leukemia. Cell. 2012;150(2):264–78.
7. Ellis P, Moore L, Sanders MA, Butler TM, Brunner SF, Lee-Six H, et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. Nat Protoc. 2021;16(2):841–71.
8. Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing S V, Ellis P, et al. Somatic mutation landscapes at single-molecule resolution. Nature. 2021;(November 2020).
9. Luquette LJ, Miller MB, Zhou Z, Bohrson CL, Galor A, Lodato MA, et al. Ultraspecific somatic SNV and indel detection in single neurons using primary template-directed amplification. bioRxiv. 2021 Jan 1;2021.04.30.442032.
10. Xing D, Tan L, Chang CH, Li H, Xie XS. Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. Proc Natl Acad Sci U S A. 2021/02/18. 2021;118(8).
11. Schon EA, DiMauro S, Hirano M. Human mitochondrial DNA: roles of inherited and somatic mutations. Nat Rev Genet. 2012 Dec;13(12):878–90.
12. Alston CL, Rocha MC, Lax NZ, Turnbull DM, Taylor RW. The genetics and pathology of mitochondrial disease. J Pathol. 2016/11/02. 2017 Jan;241(2):236–50.
13. Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. Nat Rev Genet [Internet]. 2005;6(5):389–402. Available from: https://doi.org/10.1038/nrg1606
14. Greaves LC, Reeve AK, Taylor RW, Turnbull DM. Mitochondrial DNA and disease. J Pathol. 2012 Jan 1;226(2):274–86.
15. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538:260–4.
16. Franco I, Johansson A, Olsson K, Vrtacnik P, Lundin P, Helgadottir HT, et al. Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. Nat Commun. 2018;9(1):800.
17. Yoshida K, Gowers KHC, Lee-Six H, Chandrasekharan DP, Coorens T, Maughan EF, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature. 2020;578(7794):266–72.
18. Grossmann S, Hooks Y, Wilson L, Moore L, O'Neill L, Martincorena I, et al. Development, maturation, and maintenance of human prostate inferred from somatic mutations. Cell Stem Cell. 2021;28(7):1262-1274.e5.
19. Mitchell E, Chapman MS, Williams N, Dawson K, Mende N, Calderbank EF, et al. Clonal dynamics of haematopoiesis across the human lifespan. bioRxiv. 2021;
20. Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, et al. The mutational landscape of human somatic and germline cells. Nature. 2021;597(7876):381–6.
21. Park S, Mali NM, Kim R, Choi J-W, Lee J, Lim J, et al. Clonal dynamics in early human embryogenesis inferred from somatic mutation. Nature. 2021;597(7876):393–7.
22. Lee-Six H, Øbro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature. 2018;561(7724):473–8.
23. Osorio FG, Rosendahl Huber A, Oka R, Verheul M, Patel SH, Hasaart K, et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. 2018 Nov;25(9):2308-2316.e4.

24. Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature. 2019;574(7779):538–42.
25. Franco I, Helgadottir HT, Moggio A, Larsson M, Vrtačnik P, Johansson A, et al. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. Genome Biol. 2019;20(1).
26. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;574(7779):532–7.
27. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019;565(7739):312–7.
28. Lawson ARJ, Abascal F, Coorens THH, Hooks Y, O'Neill L, Latimer C, et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. Science (80-). 2020;370(6512):75–82.
29. Moore L, Leongamornlert D, Coorens THH, Sanders MA, Ellis P, Dentro SC, et al. The mutational landscape of normal human endometrial epithelium. Nature. 2020;580(7805):640–6.
30. Rahbari R, Consortium UK, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, et al. Timing, rates and spectra of human germline mutation. Nat Genet. 2016;48(2):126–33.
31. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017;549(7673):519–22.
32. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. 2013;3:246–59.
33. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91.
34. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci. 2004 Mar 23;101(12):4164 LP – 4169.
35. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149(5):979–93.
36. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
37. Huang X, Wojtowicz D, Przytycka TM. Detecting presence of mutational signatures in cancer with confidence. Bioinformatics. 2018;34(2):330–7.
38. Wang S, Li H, Song M, Tao Z, Wu T, He Z, et al. Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. PLOS Genet. 2021 May 4;17(5):e1009557.
39. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578(7793):112–21.
40. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534(7605):1–20.
41. Christensen S, Van der Roest B, Besselink N, Janssen R, Boymans S, Martens JWM, et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun. 2019;10(1):4571.
42. Pleguezuelos-Manzano C, Puschhof J, Rosendahl Huber A, van Hoeck A, Wood HM, Nomburg J, et al. Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature. 2020;580:269–73.
43. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177:821-836.e16.
44. Degasperi A, Amarante TD, Czarnecki J, Shooter S, Zou X, Glodzik D, et al. A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. Nat cancer. 2020;1:249–63.
45. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 2018;
46. Aitken SJ, Anderson CJ, Connor F, Pich O, Sundaram V, Feig C, et al. Pervasive lesion segregation

shapes cancer genome evolution. Nature. 2020;583:265–70.
47. Petljak M, Chu K, Dananberg A, Bergstrom EN, Morgen P von, Alexandrov LB, et al. The APOBEC3A deaminase drives episodic mutagenesis in cancer cells. bioRxiv. 2021 Jan 1;2021.02.14.431145.
48. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet. 2015;47:1402–7.
49. Li R, Di L, Li J, Fan W, Liu Y, Guo W, et al. A body map of somatic mutagenesis in morphologically normal human tissues. Nature. 2021/08/27. 2021;597(7876):398–403.
50. Tang J, Fewings E, Chang D, Zeng H, Liu S, Jorapur A, et al. The genomic landscapes of individual melanocytes from human skin. Nature. 2020;586(7830):600–5.
51. Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell. 2019/03/09. 2019;176(6):1282-1294.e20.
52. Wyles SP, Brandt EB, Nelson TJ. Stem cells: the pursuit of genomic stability. Int J Mol Sci. 2014;15(11):20948–67.
53. Brazhnik K, Sun S, Alani O, Kinkhabwala M, Wolkoff AW, Maslov AY, et al. Single-cell analysis reveals different age-related somatic mutation profiles between stem and differentiated cells in human liver. Sci Adv. 2020;6(5):eaax2659.
54. Zhang L, Dong X, Lee M, Maslov AY, Wang T, Vijg J. Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. Proc Natl Acad Sci U S A. 2019;116(18):9014–9.
55. Machado H, Mitchell E, Obro N, Kubler K, Davies M, Maura F, et al. Genome-wide mutational signatures of immunological diversification in normal lymphocytes. bioRxiv. 2021;
56. Lodato MA, Rodin RE, Bohrson CL, Coulter ME, Barton AR, Kwon M, et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science (80-). 2018;
57. Meldrum C, Doyle MA, Tothill RW. Next-generation sequencing for cancer diagnostics: a practical perspective. Clin Biochem Rev. 2011 Nov;32(4):177–95.
58. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun [Internet]. 2015;6(1):10001. Available from: https://doi.org/10.1038/ncomms10001
59. Behjati S, Huch M, Van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature. 2014;513(7518):422–5.
60. Ju YS, Martincorena I, Gerstung M, Petljak M, Alexandrov LB, Rahbari R, et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature. 2017;543(7647):714–8.
61. Hasaart KAL, Manders F, van der Hoorn M-L, Verheul M, Poplonski T, Kuijk E, et al. Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis. Sci Rep. 2020;10(1):12991.
62. Spencer Chapman M, Ranzoni AM, Myers B, Williams N, Coorens THH, Mitchell E, et al. Lineage tracing of human development through somatic mutations. Nature. 2021;595(7865):85–90.
63. Bae T, Tomasini L, Mariani J, Zhou B, Roychowdhury T, Franjic D, et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. Science (80-). 2018;359(6375)
64. Kuijk E, Blokzijl F, Jager M, Besselink N, Boymans S, Chuva de Sousa Lopes SM, et al. Early divergence of mutational processes in human fetal tissues. Sci Adv. 2019;5(5):eaaw1271.
65. Baron CS, van Oudenaarden A. Unravelling cellular relationships during development and regeneration using genetic lineage tracing. Nat Rev Mol Cell Biol. 2019;20(12):753–65.
66. Ju YS, Martincorena I, Gerstung M, Petljak M, Alexandrov LB, Rahbari R, et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature. 2017;543(7647):714–8.
67. Vanneste E, Voet T, Le Caignec C, Ampe M, Konings P, Melotte C, et al. Chromosome instability is common in human cleavage-stage embryos. Nat Med. 2009;15(5):577–83.
68. Palmer N, Kaldis P. Regulation of the Embryonic Cell Cycle During Mammalian Preimplantation Development. Curr Top Dev Biol. 2016;1–53.
69. Schulz KN, Harrison MM. Mechanisms regulating zygotic genome activation. Nat Rev Genet.

2019;20(4):221–34.
70. Colaco S, Sakkas D. Paternal factors contributing to embryo quality. J Assist Reprod Genet. 2018/09/13. 2018;35(11):1953–68.
71. McCoy RC. Mosaicism in Preimplantation Human Embryos: When Chromosomal Abnormalities Are the Norm. Trends Genet. 2017;33(7):448–63.
72. Vazquez-Diez C, FitzHarris G. Causes and consequences of chromosome segregation error in preimplantation embryos. Reproduction. 2017/11/08. 2018;155(1):R63–76.
73. Pilati C, Shinde J, Alexandrov LB, Assié G, André T, Hélias-Rodzewicz Z, et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. J Pathol. 2017;242(March):10–5.
74. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646–74.
75. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. Science (80-). 2019;366(6465).
76. Kakiuchi N, Ogawa S. Clonal expansion in non-cancer tissues. Nat Rev Cancer. 2021;21(4):239–56.
77. Mustjoki S, Young NS. Somatic Mutations in “Benign” Disease. Reply. N Engl J Med. 2021;385(11):e34.
78. Vijg J, Dong X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. Cell. 2020;182(1):12–23.
79. Cagan A, Baez-Ortega A, Brzozowska N, Abascal F, Coorens THH, Sanders MA, et al. Somatic mutation rates scale with lifespan across mammals. bioRxiv. 2021;
80. Tiwari V, Wilson 3rd DM. DNA Damage and Associated DNA Repair Defects in Disease and Premature Aging. Am J Hum Genet. 2019;105(2):237–57.
81. Robinson PS, Coorens THH, Palles C, Mitchell E, Abascal F, Olafsson S, et al. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. Nat Genet. 2021;53(10):1434–42.
82. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. Nat Rev Genet. 2013;14(5):307–20.
83. Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. Somatic mosaicism: implications for disease and transmission genetics. Trends Genet. 2015;31(7):382–92.
84. Fernández LC, Torres M, Real FX. Somatic mosaicism: on the road to cancer. Nat Rev Cancer. 2016;16(1):43–55.
85. Priest JR, Gawad C, Kahlig KM, Yu JK, O’Hara T, Boyle PM, et al. Early somatic mosaicism is a rare cause of long-QT syndrome. Proc Natl Acad Sci U S A. 2016;113(41):11555–60.
86. Poduri A, Evrony GD, Cai X, Walsh CA. Somatic Mutation, Genomic Variation, and Neurological Disease. Science (80-). 2013;341(6141):1237758.
87. McConnell MJ, Moran J V, Abyzov A, Akbarian S, Bae T, Cortes-Ciriano I, et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science (80-). 2017;356(6336).
88. Paquola ACM, Erwin JA, Gage FH. Insights into the role of somatic mosaicism in the brain. Curr Opin Syst Biol. 2017;1:90–4.
89. Hodges H, Fealko C, Soares N. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. Transl Pediatr. 2020;9(Suppl 1):S55–65.
90. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. 2020;578(7793):122–8.
91. Williams N, Lee J, Moore L, Joanna Baxter E, Hewinson J, Dawson KJ, et al. Phylogenetic reconstruction of myeloproliferative neoplasm reveals very early origins and lifelong evolution. bioRxiv. 2020;
92. Filbin M, Monje M. Developmental origins and emerging therapeutic opportunities for childhood cancer. Nat Med. 2019;25(3):367–76.
93. Greaves M. A causal mechanism for childhood acute lymphoblastic leukaemia. Nat Rev Cancer.

2018;18(8):471–84.
94. Hardy K, Hardy PJ. 1(st) trimester miscarriage: four decades of study. Transl Pediatr. 2015;4(2):189–200.
95. Yang X, Breuss MW, Xu X, Antaki D, James KN, Stanley V, et al. Developmental and temporal characteristics of clonal sperm mosaicism. Cell. 2021;184(18):4772-4783.e15.
96. Veltman JA, Brunner HG. De novo mutations in human genetic disease. Nat Rev Genet. 2012;13(8):565–75.
97. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433–8.
98. Iakoucheva LM, Muotri AR, Sebat J. Getting to the Cores of Autism. Cell. 2019;178(6):1287–98.
99. Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290(5806):457–65.
100. Satoh M, Kuroiwa T. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Exp Cell Res [Internet]. 1991;196(1):137–40. Available from: https://www.sciencedirect.com/science/article/pii/0014482791904679
101. García-Rodríguez LJBT-M in CB. Appendix 1. Basic Properties of Mitochondria. In: Mitochondria, 2nd Edition. Academic Press; 2007. p. 809–12.
102. Stewart JB, Chinnery PF. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat Rev Genet. 2015;16(9):530–42.

Chapter 2

Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis

Karlijn A. L. Hasaart[1,†], Freek Manders[1,†], Marie-Louise van der Hoorn[2], Mark Verheul[1], Tomasz Poplonski[3], Ewart Kuijk[4], Susana M. Chuva de Sousa Lopes,[5] and Ruben van Boxtel[1,*]

[1]Princess Máxima Center for Pediatric Oncology and Oncode Institute, Heidelberglaan 25, 3584CS Utrecht, The Netherlands
[2]Leiden University Medical Center, 2333 ZC, Leiden, The Netherlands
[3]Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584CS Utrecht, The Netherlands
[4]Center for Molecular Medicine, University Medical Center Utrecht and Oncode Institute, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands
[5]Department of Anatomy and Embryology, Leiden University Medical Center, 2333 ZC, Leiden, The Netherlands
[†]These authors contributed equally
[*]Corresponding author: R.vanBoxtel@prinsesmaximacentrum.nl

Scientific Reports, 2020

Abstract

Children show a higher incidence of leukemia compared to young adolescents, yet their cells have less age-related (oncogenic) somatic mutations. Newborns with Down syndrome have an even higher risk of developing leukemia, which is thought to be driven by mutations that accumulate during fetal development. To characterize mutation accumulation in individual stem and progenitor cells of Down syndrome and karyotypically normal fetuses, we clonally expanded single cells and performed whole-genome sequencing. We found a higher mutation rate in haematopoietic stem and progenitor cells during fetal development compared to the post-infant rate. In fetal trisomy 21 cells the number of somatic mutations is even further increased, which was already apparent during the first cell divisions of embryogenesis before gastrulation. The number and types of mutations in fetal trisomy 21 haematopoietic stem and progenitor cells were similar to those in Down syndrome-associated myeloid preleukemia and could be attributed to mutational processes that were active during normal fetal haematopoiesis. Finally, we found that the contribution of early embryonic cells to human fetal tissues can vary considerably between individuals. The increased mutation rates found in this study, may contribute to the increased risk of leukemia early during life and the higher incidence of leukemia in Down syndrome.

Introduction

The initiation and progression of cancer is thought to result from somatic clonal evolution in human tissues[1]. DNA mutations promote heritable phenotypic diversity in cell populations, which provides the substrate for context-dependent selection forces. Oncogenic mutations allow cells to become independent of external growth factors, or insensitive to intrinsic inhibitory signals, which in the correct context can promote uncontrolled clonal expansion and eventually cancer. For adult cancers, the acquisition of oncogenic mutations is thought to be rate limiting for tumor initiation, providing an explanation why aging is the biggest risk factor for developing cancer[2,3]. Indeed, somatic mutations accumulate gradually throughout human life[4,5]. However, children can also develop cancer. In fact, for some cancers, such as leukemia, the incidence is higher in children compared to adolescents, even though their young cells have less age-related (oncogenic) mutations[6]. The mutations driving pediatric leukemia are thought to be acquired during fetal development[7]; however, the rate and patterns of mutation accumulation in haematopoietic stem and progenitor cells (HSPCs) during fetal development are currently not known.

Newborns with Down syndrome (DS) provide an opportunity to better understand the molecular mechanisms underlying pediatric leukemogenesis, because DS children show a substantially elevated risk of developing leukemia during their first years of life8. Children with DS have a 500 fold higher risk of developing acute megakaryoblastic leukemia (DS-AMKL) compared to the general population[8,9]. DS-AMKL is often preceded by DS-associated myeloid preleukemia, which is observed in 5-10% of all DS newborns and usually spontaneously disappears within the first 3-4 months after birth[9]. However, even when spontaneous regression is achieved, approximately 20-30% of all DS-associated myeloid preleukemia patients will develop DS-AMKL[10]. This suggests that an extra copy of chromosome 21 can act as a genetic driver of cancer, but that additional oncogenic driver mutations are required[11,12]. In line with this, DS-associated myeloid preleukemia is characterized by somatic mutations in GATA1, which cause a N-terminally truncated protein[13]. These GATA1 mutations are acquired during fetal development and are sufficient for the development of DS-associated myeloid preleukemia[13,14]. Remarkably, it has been reported that in some DS-associated myeloid preleukemia patients several independent clones exist, which are characterized by distinct GATA1 mutations[12]. This observation suggests that the HSPCs in the fetal liver of DS fetuses might be subjected to high levels of mutagenesis. Previously, it has been shown that aneuploidy in yeast results in genomic instability[15]. However, it is not known if an aneuploidy of chromosome 21 causes an increase in somatic mutation load in cells of human trisomy 21 (T21) fetuses. To compare the somatic mutation rates and patterns during normal and T21 fetal development, we studied mutation accumulation in single HSPCs and intestinal stem cells (ISCs) of fetuses with a normal karyotype and of fetuses with T21. We found an increased somatic mutation rate in fetal HSPCs and even higher somatic mutation numbers in cells of T21 fetuses. Moreover, we found that somatic mutations in DS-associated preleukemia can be explained by mutational processes, which are normally active in normal and T21 fetal haematopoiesis. Second, we showed that the contribution of developmental lineage branches to fetal tissues can be symmetric as well as asymmetric. This observation indicates that the contribution of developmental lineage branches to tissues can vary between fetuses, independent of T21.

Results

Mutation accumulation during human fetal haematopoiesis

Cataloguing somatic mutations in physiologically normal cells is technically challenging due to the polyclonal nature of healthy tissues and the high error rate of single cell sequencing techniques[16]. Previously, we have developed a method to characterize somatic mutations in single cells using clonal cultures of primary human stem cells of various tissues[17], including adult HSPCs[4]. Here, we applied a similar approach to catalogue somatic mutation in fetal HSPCs as well as donor-matched ISCs (Fig. 1). We included 9 independent human fetuses gestational age (GA) week 12-17) (Supplementary Table S1 online). Four of these fetuses had a constitutive T21 and five of these fetuses were karyotypically normal (D21) (Supplementary Table S1 online). We isolated HSPCs (CD34+, lineage -) from liver and bone marrow (Supplementary Fig. S1 online) and clonally expanded these cells for 3-4 weeks in culture to obtain sufficient DNA for whole-genome sequencing (WGS)[18]. Moreover, we clonally expanded ISCs of the same fetus into organoid cultures for 6-7 weeks and performed WGS. From each fetus, we sequenced DNA from bulk skin or intestine to control for germline variants (see Methods). This approach allowed us to obtain all the mutations that were present in the originally expanded fetal stem and progenitor cells and which were acquired *in vivo*[17,18]. Mutations that accumulated during the *in vitro* expansion could be excluded based on their low variant allele frequency (Supplementary Fig. S2 online), as not all the cells in the clonal culture share these mutations in contrast to the *in vivo* acquired mutations.

In total, we observed 740 base substitutions and 42 indels in 17 clonal D21 HSPCs and 11 clonal D21 ISC cultures, which were obtained from 5 independent fetuses (Fig 2a, Supplementary Table S2 online). In addition, we found 873 base substitutions and 41 indels in 14 clonal T21 HSPCs and 9 clonal T21 ISC cultures obtained from 4 independent fetuses (Supplementary Table S2 online). We did not observe any larger structural variants or chromosomal aberrations (see Methods). Almost all somatic mutations were located in introns. In total we found 11 somatic mutations located in exons in D21 fetal stem and progenitor cells and 8 in T21 fetal stem and progenitor cells, none of which we considered to be drivers (Supplementary table S3 online) (see Methods). Moreover, we did not observe a mutation in GATA1 in any of the fetal stem and progenitor cells, suggesting that there is no myeloid preleukemia clone present. There was no significant difference in the types of somatic exonic mutations between D21 and T21 fetal stem and progenitor cells (p = 0.578, chi-squared test) (Fig. 2b). In addition, we compared our data to genome-wide mutation catalogues observed in D21 post-infant HSPCs and D21 post-infant ISCs obtained from our previous studies[4,5]. We calculated the somatic mutation rate of HSPCs and ISCs during fetal development and after birth by dividing the number of somatic mutations by the age (in years) of the fetus or donor since conception. We observed an annual somatic mutation rate in D21 fetal HSPCs of approximately 100 base substitutions per year (95% confidence interval: 88 – 113), which is 5.8 times higher compared to the rate observed in D21 post-infant HSPCs (p = 1.231 x 10-5, linear mixed-effects model) (Fig. 2a). We also observed a higher mutation rate in D21 fetal ISCs compared to D21 post-infant ISCs (p = 0.00153, linear mixed-effects model) (Fig. 2a), which is line with a previous study that catalogued somatic mutations in D21 fetal ISCs[19].

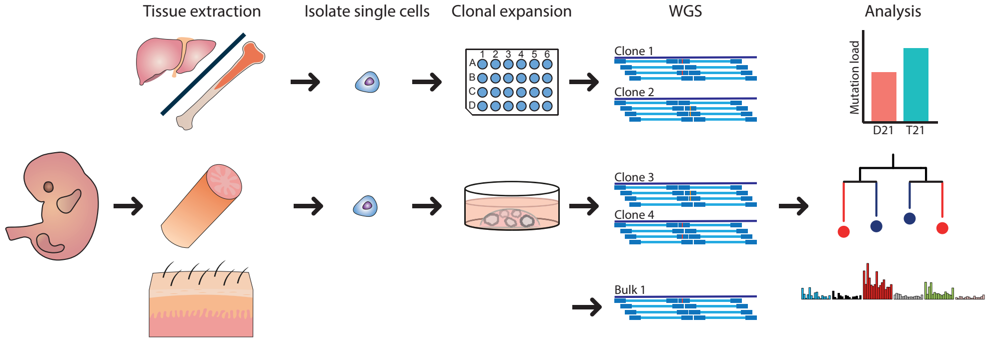


Fig. 1: Characterizing somatic mutations in single fetal haematopoietic stem and progenitor cells (HSPCs) and fetal intestinal stem cells (ISCs).
Experimental strategy for characterizing somatic mutations in single cells of disomy 21 (D21) and trisomy (T21) fetuses. HSPCs and ISC were clonally expanded to obtain sufficient DNA for whole-genome sequencing (WGS). DNA from bulk skin or intestine was used as reference to control for germline variants. After characterizing the somatic mutations in single cells, the somatic mutation load between D21 and T21 fetal cells was compared. In addition, signature analysis and phylogenetic lineage tree analyses were performed.

Increased mutation load during fetal development in Down syndrome

T21 stem and progenitor cells of T21 fetuses accumulated about 34 (95% confidence interval: 6 – 62) extra somatic base pair substitutions mutations per cell compared to D21 stem and progenitor cells during fetal development (*p* = 0.0239, linear mixed-effects model) (Fig. 2c,d; Supplementary Fig. S3, S4 online). Of note, this increase was not restricted to the haematopoietic system. We validated that the difference in mutation load between T21 and D21 fetal stem and progenitor cells was not dependent on one or even two single data points (Supplementary Fig. S4 online), underlining the robustness of our finding. Interestingly, T21 fetal stem and progenitor cells also showed an increased variance in somatic mutation load compared to D21 fetal cells (*p* = 2.1 x 10-8, likelihood-ratio test, LR: 31) with some cells showing 2 – 3 times higher mutation load than age-matched D21 fetal cells (Fig. 2c). Of these, one HSPC of a GA week 14.5 T21 fetus showed a significantly higher number of somatic mutations than expected compared to other T21 fetal stem and progenitor cells

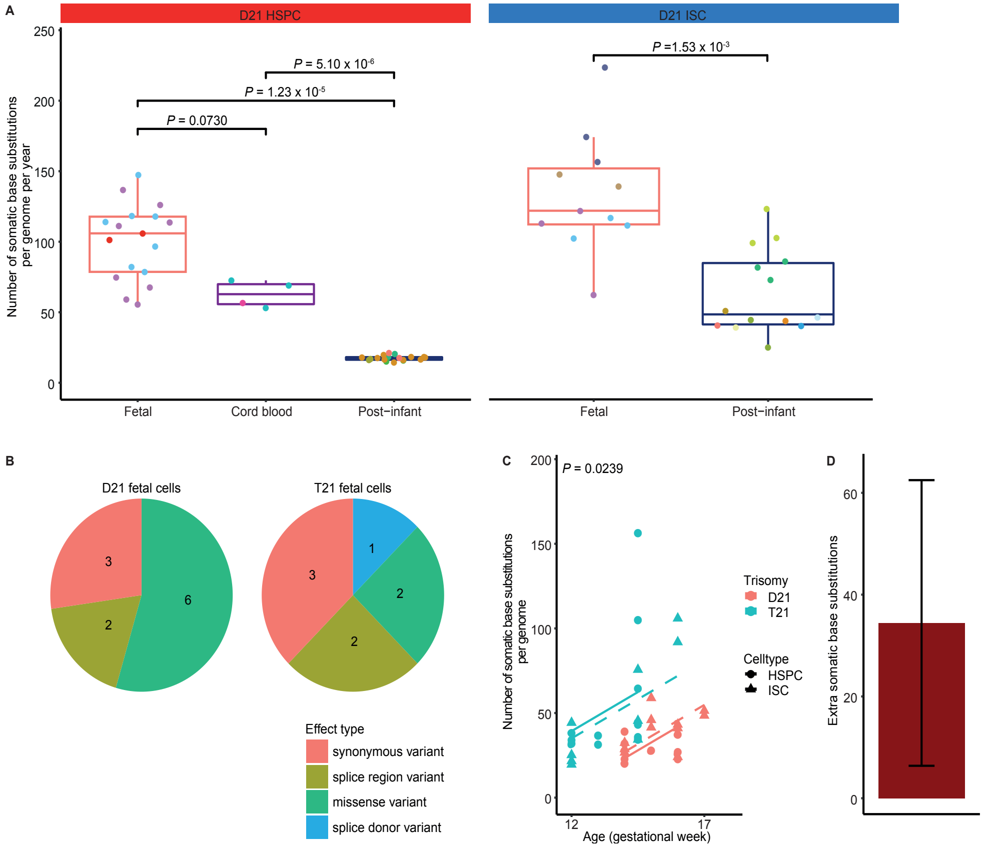


Fig. 2: Accumulation of somatic base pair substitutions in haematopoietic stem and progenitor cells (HSPCs) and intestinal stem cells (ISCs) during human fetal development and after birth.
a Comparison of the number of autosomal somatic base substitutions per genome per year between D21 HSPCs (D21 fetal: 17 clones; 3 donors, Cord blood: 4 clones; 2 donors, Post-infant: 18 clones; 6 donors) and D21 ISCs (D21 fetal: 11 clones; 4 donors, Post-infant: 14 clones; 9 donors) of fetuses, cord blood and post infant (linear mixed-effects model). Points with the same color indicate single cells from the same subject. **b** Pie charts showing the number of somatic mutations for different types of exonic mutations in D21 and T21 fetal stem and progenitor cells (D21 fetal: 28 clones; 5 donors, T21 fetal: 23 clones; 4 donors). **c** The number of somatic base substitutions per genome plotted against the donor age (D21 fetal: 28 clones; 5 donors, T21 fetal: 23 clones; 4 donors). Dashed line: ISC, full line: HSPC. *p*-value shows the difference between T21 and D21 fetal stem and progenitor cells. (linear mixed-effects model, two-tailed t-test). d Extra somatic base substitutions per genome in T21 fetal stem and progenitor cells. Error bars represent 95% confidence intervals.

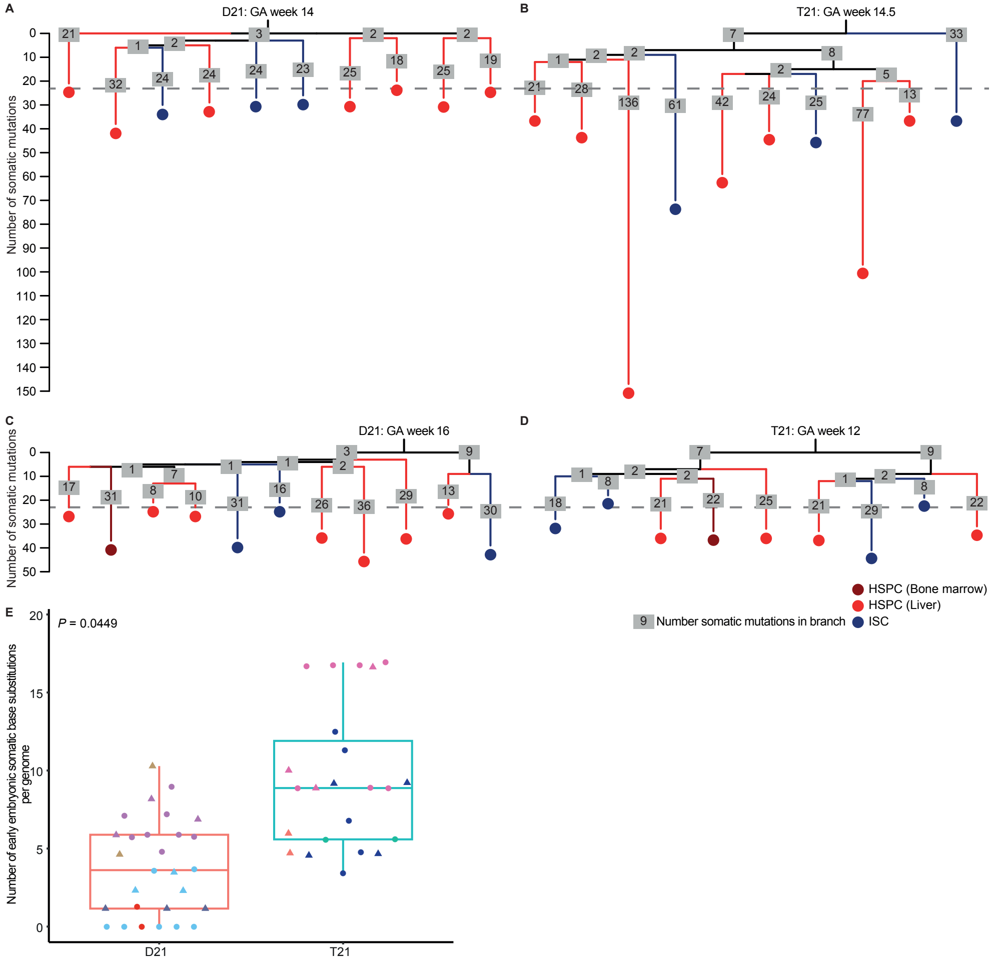


Fig. 3: Phylogenetic lineage trees of disomy 21 (D21) and trisomy 21 (T21) fetuses.
a Lineage trees of a gestational age (GA) week 14 D21 fetus, **b** gestational age week 14,5 T21 fetus, **c** gestational age week 16 D21 fetus and **d** gestational age week 12 T21 fetus. Each tip represents a single clonally expanded cell. The length of the branches indicates the number of somatic mutations in that branch of the tree. The number of somatic mutations in each branch are shown in grey boxes. **e** Comparison of the number of somatic base substitutions per genome between D21 and T21 fetal stem and progenitor cells, that occurred early in the development of the fetus (D21 fetal: 28 clones; 5 donors, T21 fetal: 23 clones; 4 donors). Circle: Haematopoietic stem and progenitor cells, Triangle: intestinal stem cells. Points with the same color indicate single cells from the same subject. (linear mixed-effects model, two-tailed t-test).

(linear mixed-effects model, two-sided outlier test, FDR = 0.049) (Fig. 2c). Finally, we observed 9 double base pair substitution (DBS) in T21 fetal stem and progenitor cells versus only 1 DBS in D21 fetal stem and progenitor cells (p = 0.0017, Wilcoxon test) (Supplementary Fig. S5 online). We did not observe a difference in the number of indels between T21 and D21 fetal stem and progenitor cells (p = 0.815, linear mixed-effects model) (Supplementary Fig. S6 online). Taken together, our results show that the presence of a constitutive T21 in T21 fetuses results in an increased number of base substitutions as well as an increased variance in mutation load between different stem and progenitor cells.

Mutation accumulation during early embryogenesis

To determine when during fetal development the difference in mutation load between T21 and D21 fetal stem and progenitor cells occurred, we used a phylogenetic analysis approach to time the occurrence of somatic mutations during development (see Methods). Somatic mutations that are shared between two fetal stem and progenitor cells reflect a historical common ancestor. The more mutations two cells share, the later during development these two cells separated from a common ancestral cells[20,21]. By assessing all the mutations that are shared between the different cells of the same fetus, we constructed developmental lineage trees for 2 D21 and 2 T21 fetuses (Fig. 3a,b,c,d). Mutations near the trunk of the developmental lineage tree are shared between endoderm-derived ISCs and mesoderm-derived HSPCs, indicating that these mutations were acquired before gastrulation. In line with this, these mutations also showed sub-clonal presence in the matching skin bulk sample, which is derived from ectoderm. We used this analysis to compare the mutation rates during early embryonic development between T21 and D21 fetuses (see Methods). We found about 6 (95% confidence interval: 0.2 – 11.7) extra somatic mutations per cell acquired during the first cell divisions in T21 fetal stem and progenitor cells compared to D21 fetal stem and progenitor cells (p = 0.0449, linear mixed-effects model) (Fig. 3e). This observation indicates that the mutation load is already increased in T21 very early after conception, before gastrulation.

Contribution of developmental lineage branches to D21 and T21 fetal tissues

We used the developmental lineage trees to study the contribution of early embryonic branches to bulk skin in each fetus. For this analysis, we compared for each fetus the median VAF in bulk skin of the somatic mutations accumulated before gastrulation between the first 2 developmental lineages branches (Fig. 4). Because all somatic mutations are accumulated during the same period of fetal development, we were able to compare fetuses of different GA. Previous studies using similar mutational analyses in adult HSPCs obtained from human donors revealed an asymmetric contribution of developmental branches to the adult haematopoietic system[4,21,22]. Also in adult mice, this asymmetric contribution of developmental branches was observed20. These observations indicate that early embryonic cells do not contribute equally to adult tissues. In line with this, we found that the phylogenetic lineage trees of a GA week 16 D21 fetus and a GA week 12 T21 fetus showed an asymmetric contribution of the first 2 detectable development lineage branches (p = 2.259 x 10-13, 1.525 x 10-13, chi-squared test) (Fig. 4c,d). However, we did not observe this asymmetric contribution in a GA week 14.5 T21 and a GA week 14 D21 fetus (Fig. 4a,b). This observation indicates that the contribution of early embryonic cells to fetal tissues can vary between fetuses, independent of T21.

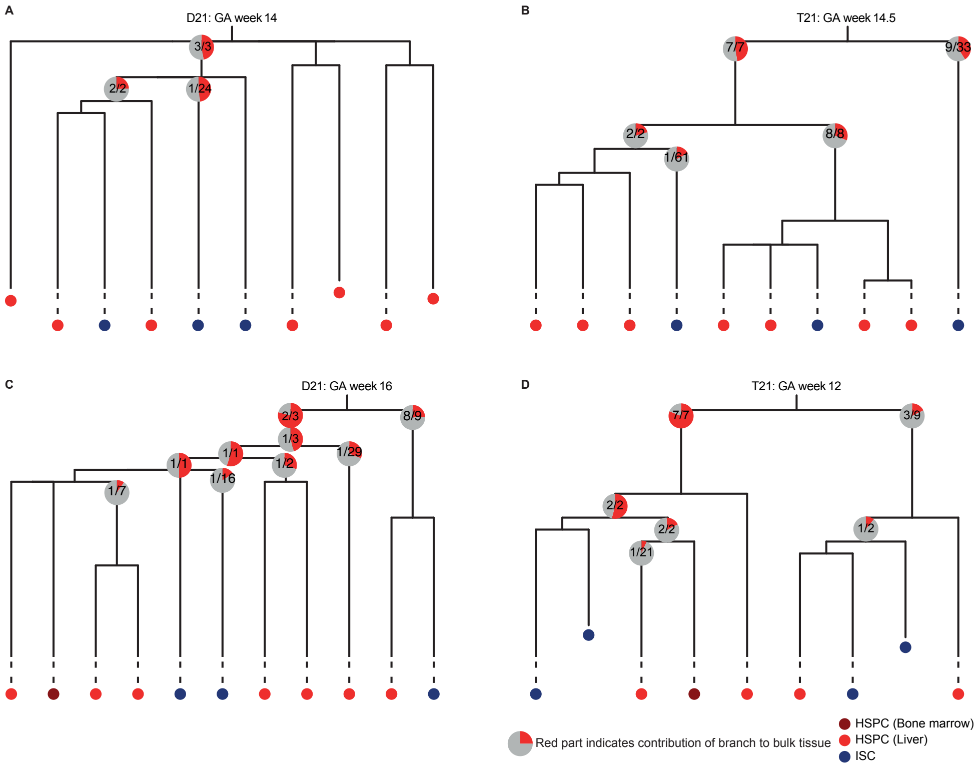


Fig. 4: Relative contribution of the developmental lineage branches to fetal skin tissue. **a** Lineage trees of a gestational age (GA) week 14 D21 fetus, **b** gestational age week 14,5 T21 fetus, **c** gestational age week 16 D21 fetus and **d** gestational age week 12 T21 fetus. Each tip represents a single clonally expanded cell. The pie charts show the median contribution of the contributing mutations in a branch to the bulk skin tissue. The grey part of the pie chart indicates the total skin tissue, while the red part shows the contribution of a single branch to the skin tissue. The text in the pie charts shows how many of the mutations in that branch contributed to the skin tissue. Mutations not contributing to the skin tissue at all, are not used to calculate the median. Multiple pie charts in a single branch indicate that the mutations in that branch occurred during different cell divisions.

Activity of mutational processes in post-infant, D21 and T21 fetal haematopoiesis

To identify the processes underlying somatic mutation accumulation in fetal HSPCs, we determined the relative contribution of previously defined mutational signatures to the observed mutation spectra (see Methods)[23,24]. The mutation spectra between D21 fetal HSPCs and D21 post-infant HSPCs were significantly different (chi-squared test, p = 5.0 x 10-4) (Fig. 5a; Supplementary Fig. S7 and S8 online), which in part can be explained by a higher relative contribution of single base substitution signature 1 (SBS1) in fetal compared to D21 post-infant HSPCs (P < 5.0 x 10-3, permutation test) (Fig 5b; Supplementary Fig. S8 online). The underlying mechanism of SBS1 is thought to be the spontaneous deamination of methylated cytosines, which likely reflects a cell cycle-dependent mutational clock[5,25]. Moreover, the relative contribution of the recently defined HSPC-specific mutational signature [4,22,24] was less present in D21 fetal HSPCs, whereas it is predominant in D21 post-infant HSPCs4 (P < 5.0 x 10-3, permutation test) (Fig. 5b; Supplementary Fig. S8 online). In contrast, we did not find any difference between the mutation spectra of D21 fetal ISCs and D21 post-infant ISCs (p = 0.460, chi-squared test) (Supplementary Fig. S9 online), which is in line with a previous study[19]. The mutation spectra and relative contribution of mutational signatures between D21 and T21 fetal cells did not differ for HSPCs and ISCs. This indicates that the same mutational processes can explain the somatic mutations in D21 and T21 fetal stem and progenitor cells (Fig. 5a,b; Supplementary Fig. S9 online). Of note, the T21 fetal HSPC with a significantly higher mutation load compared to other T21 fetal cells did show contribution of an additional signature SBS18 (Fig. 5c), which has previously been associated with oxidative stress-induced mutagenesis[26]. Interestingly, an increase in the generation of radical oxygen species has been reported in T21 neurons, suggesting that ROS is preserved in several cell types in T21[27]. Our findings indicate that the increased mutation load in T21 fetal stem and progenitor cells is mostly caused by processes that are active during normal fetal development, suggesting that there is more activity of these mutational processes in T21 fetal stem and progenitor cells. However, the increased variance in mutation load might be explained by the activity of additional processes, such as oxidative stress-induced mutagenesis.

T21 HSPCs display similar mutation load and patterns as DS-associated myeloid preleukemia

We compared the somatic mutation load of preleukemic blast cells from 6 independent DS-associated myeloid preleukemia patients[11] with those observed in the T21 fetal HSPCs and found similar numbers of mutations (p = 0.643, linear mixed-effects model) (Fig. 6a). As mutation accumulation in normal stem cells acts as a molecular clock[4,5], our findings suggest that the DS-associated myeloid preleukemias arose

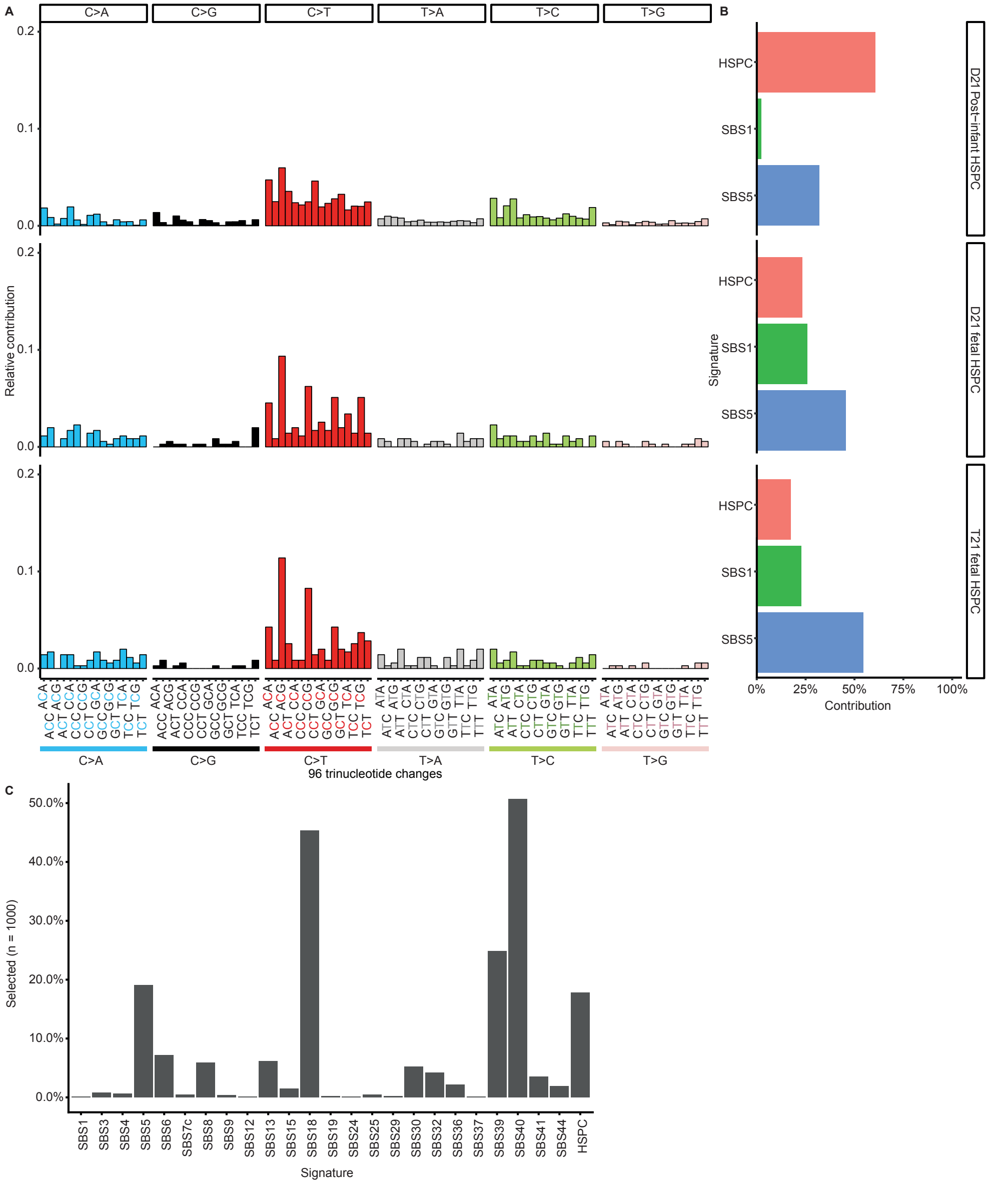


Fig. 5: Somatic mutation patterns of disomy 21 (D21) and Trisomy 21 (T21) haematopoietic stem and progenitor cells (HSPCs).
a Spectra of somatic point substitutions. The substitutions are pooled per category. (D21 Post-infant HSPC: n = 10924; 18 clones; 5 donors, D21 fetal HSPC: n = 353; 17 clones; 3 donors, T21 fetal HSPC: n = 351; 13 clones; 3 donors). **b** The relative contribution of each mutational signature to the spectra of point substitutions. **c** Bar plot depicting how often each signature was selected during a bootstrapped (1000 iterations) signature selection process for the T21 fetal HSPC with extremely high somatic mutation load.

during the period of fetal development that we assessed. In all preleukemic blast cells of DS-associated myeloid preleukemia patients we observed GATA1 mutations

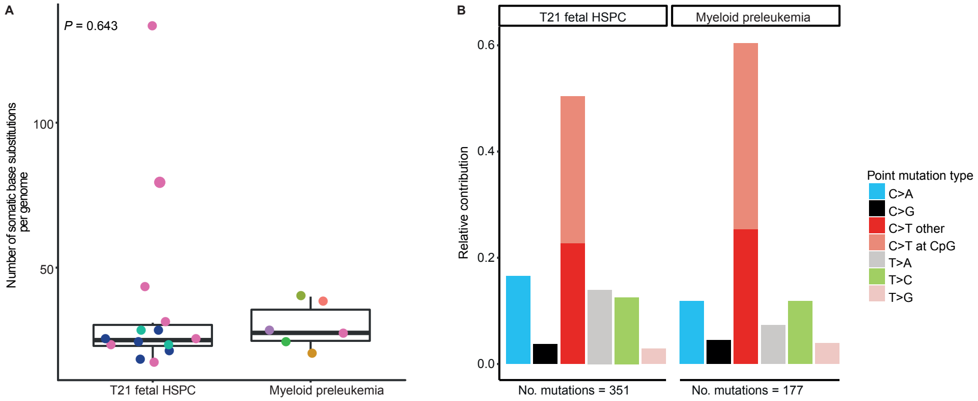


Fig. 6: Somatic mutation patterns of preleukemic bulk blast cells from DS-associated myeloid preleukemia patients.
a Comparison of the number of autosomal somatic base substitutions per genome per year for T21 fetal haematopoietic stem and progenitor cells (14 clones; 3 donors) and DS-associated myeloid preleukemia (6 donors). Points with the same color indicate single cells from the same subject. (linear mixed-effects model, two-tailed t-test). **b** 7-Spectrum of somatic base substitutions. The total number of somatic base substitutions is indicated.

(Supplementary table S3 online). However, no additional clonal cancer driver mutations were identified (Supplementary table S3). Moreover, we found no difference in the mutation spectra between T21 fetal HSPCs and DS-associated myeloid preleukemia, suggesting that SBS1, SBS5 and HSPC can explain the clonal somatic mutations in preleukemic blast cells of DS-associated myeloid preleukemia patients ($p = 0.164$, chi-squared test) (Fig. 6b, Supplementary Fig. S10 online). This observation indicates that no additional mutational processes are required to explain the somatic mutations in DS-associated myeloid preleukemia, besides those already active during normal fetal haematopoiesis.

Discussion

In the present study we characterized mutation accumulation in individual HSPCs and ISCs of D21 and T21 fetuses. Several reports have demonstrated that T21 perturbs fetal haematopoiesis, which is explained by an imbalanced expression of genes involved in haematopoietic development[28,29,30]. Nonetheless, additional cancer driver mutations are needed for leukemic development11,12, suggesting that the HSPCs in the fetal liver of T21 fetuses are subjected to increased mutagenesis.

Our findings show that the somatic mutation rate in D21 fetal HSPCs is increased during normal fetal development compared to the post-infant mutation rate. Moreover, we found that the somatic mutation spectra vary between D21 fetal HSPCs and D21 post-infant HSPCs, indicating that HSPCs are exposed to different mutational processes during fetal development. Indeed, we found that the HSPC-specific signature is predominant in D21 post-infant HSPCs, while it is less present in D21 fetal HSPCs. This observation might reflect that HSPCs reside in a different niche during fetal development. In line with other studies, we suggest that HSPCs are more protected in the adult bone marrow niche[31], because HSPCs are highly proliferative during fetal development in the liver and become quiescent after they have migrated to the bone marrow[32]. In line with this, D21 post-infant HSPCs have relatively less contribution of mutational signature SBS1, which likely reflects a cell cycle-dependent mutational clock[25]. This mutational process is predominantly active during fetal hematopoiesis and can explain the increased somatic mutation rate in D21 fetal HSPCs compared to D21 post-infant HSPCs. This suggests that the increased activity of SBS1 in D21 fetal HSPCs may contribute to the relatively higher incidence of leukemia in young children compared to young adults, since an increased mutation rate increases the chance to acquire a cancer driver mutation.

Moreover, fetal cells of T21 individuals, who are at risk of developing leukemia, show an even higher somatic mutation load, which was apparent before gastrulation. The increased mutation load in T21 fetal HSPCs was mostly caused by processes, which are active during normal fetal haematopoiesis. In addition, we showed that these mutational processes are sufficient to explain the somatic mutations in DS-associated myeloid preleukemia.

Previously, an association was found between aneuploidy and an increased mutation rate in several cancers[33]. Here, we have shown that an aneuploidy of chromosome 21 in human cells results in an increased somatic mutation load, indicating that the aneuploidy might be the first hit in these cancers. Moreover, it has been shown that aneuploid yeast strains show a mutator phenotype, which is suggested to be caused by deficient DNA repair in these strains[15]. Our study shows a similar phenotype in T21 stem and progenitor cells of T21 fetuses, independent of cell type. In line with this, several studies have reported that T21 cells of DS individuals have a decreased expression of various DNA repair genes[34,35]. This raises the hypothesis that the increased mutation load in T21 fetal stem and progenitor cells is caused by deficient DNA repair.

Many clinical features associated with DS are highly variable among individuals, such as cognitive impairment, the occurrence of heart defects as well as the development of leukemia[36]. At a molecular level, we also observed a higher variance in somatic mutation load between T21 fetal stem and progenitor cells. Intriguingly, enhanced cell-cycle and gene expression variability was previously shown in aneuploid yeast strains[37], suggesting that also at a molecular level increased variance is a common characteristic among aneuploidies.

The increased somatic mutation rate in T21 fetal HSPCs will increase the chance to acquire an oncogenic driver mutation. However, the 500 times higher incidence of AMKL in DS children cannot solely be explained by the differences we observed in mutation accumulation between D21 and T21 fetal cells8. One T21 HSPC of a T21 fetus was an outlier and showed an extremely high somatic mutation load, which can partly be explained by SBS18. However, the somatic mutations observed in DS-associated myeloid preleukemia did not show contribution of SBS18. This observation indicates that the T21 HSPCs with extremely high somatic mutation load are not necessarily the cells which undergo clonal expansion and give rise to the DS-associated myeloid preleukemia. Therefore, other factors, such as cell-cell competition, selection and/or composition of the haematopoietic microenvironment are likely to also play a role in the development of leukemia in children with DS. These factors together with the observed increased somatic mutation load in T21 cells during fetal development may explain the increased risk of developing leukemia early in life for children with DS.

In addition, we used the somatic base substitutions in fetal HSPCs and ISCs to construct developmental lineage trees of four human fetuses. We found that the contribution of the developmental lineage branches to fetal tissue can vary between fetuses, which is independent of T21. Adult tissues predominantly show an asymmetric contribution of the developmental lineage branches to tissues, while we observed a symmetric as well as an asymmetric contribution of the developmental lineage branches to fetal tissues[4,21,22]. This difference may indicate that the contribution of developmental lineage branches to tissues can change during life, which might be explained by a lower death rate and/or a higher proliferation rate of cells of a developmental lineage branch later during development. Alternatively, this change might be the consequence of a bottleneck, which is also taking place during the early blastocyst-stage in the human embryo[38].

Overall, our study provides insights in the mutation accumulation and developmental lineages during early embryogenesis and fetal haematopoiesis in normal and T21 human fetal development. These findings may contribute to the increased risk of leukemia early during life and the higher incidence of leukemia in Down syndrome.

Methods

Ethical statement

The Medical Ethical Committee of the Leiden University Medical Center approved this study (P08.087). The study was performed in accordance with the guidelines and regulations of the Helsinki declaration and its later amendments or comparable ethical standards. Signed informed consent was obtained from participating women. DS-associated myeloid preleukemia samples were obtained after approval of the Biobank commission of the Princess Máxima Center for Pediatric Oncology. Signed informed consent was obtained from all parents.

Collection of human fetal material

The gestational age (GA) in weeks was determined by the measurement of first-trimester crown-rump length by ultrasonography. In this study we included 9 fetuses from GA week 12-17. The age in weeks after conception was determined by subtracting 2 weeks from the GA. Human D21 fetal material without medical indication from elective abortion material (vacuum aspiration) was collected in 0.9% NaCl (Fresenius Kabi) and stored on ice. Human T21 fetal material was obtained from pregnant women who decided to terminate pregnancy after a positive non-invasive prenatal testing (NIPT) result for trisomy 21, which was confirmed by cytogenetic confirmatory tests after invasive prenatal screening (chorionic villus sampling or an amniocentesis).

The fetal intestine, liver and long leg bones were isolated and stored in Advanced DMEM/F-12, supplemented with 1% penicillin/streptomycin, 1% GlutaMAX, and 1% HEPES 10 mM at 4 C overnight and processed next day. A piece of fetal skin was frozen down at -20 C.

Fetal liver and small intestine disassociation

Liver and intestine were disassociated into single cell solutions with collagenase digestion as follows: biopsies were minced and incubated with EBSS supplemented with 1 mg/ml collagenase type 1A (Sigma-Aldrich) and 0.1 mg/ml DNaseI (Sigma-Aldrich) for 30 minutes at 37°C, while shaking. The tissue was further digested with a pipette if needed and incubated 10 minutes more. Subsequently, cells were filtered through a 70 um Nylon cell strainer. Single cell solutions were frozen down or further processed for culturing. Clonal ISCs WGS data from fetus F100916W15 and F100916W17 were obtained from Kuijk et al 2019[19].

Clonal intestinal organoid cultures

Single intestinal cells were plated in Matrigel (Corning) droplets in limited dilution.

Cells were cultured in human ISC organoid (CHIO) medium containing: 70% Advanced DMEM/F-12 supplemented with 1% penicillin/streptomycin, 1% GlutaMAX, and 1% HEPES 10 mM, 0.5 nM WNT surrogate (produced in house), 20% RSPOI conditioned medium (produced in house), 1x B27 supplement (Thermo Fisher Scientific), 1x Primocin (Invivogen), 1 : 1000 hES cell cloning & recovery supplement (Stemgent), 10µM SB 202190 (Sigma-Aldrich), 10 mM Nicotinamide, 1.25 mM N-acetylcysteine, 0.5 µM A83-01 (Tocris Bioscience), 10 µM Rho kinase inhibitor (Abmole), 10% noggin conditioned medium (produced in house) and 50 ng/ml hEGF (PeproTech). ISC cultures from fetus E080416, F100916W16 and F100916W17 were cultured in CHIO medium with 50% WNT conditioned medium and 100 ng/ml noggin (preprotech) as described before[19]. After 2-3 days small organoids appeared and medium was changed to human CHIO medium without hES cell cloning & recovery supplement. Clonal ISC cultures were derived by picking single organoids. Clonal organoid cultures were cultured in human CHIO medium without hES cell cloning & recovery and Rho kinase inhibitor. The cultures were expanded for 6-7 weeks until there was sufficient material for whole-genome sequencing.

Isolation and culture of haematopoietic stem and progenitor cells
Mononuclear cells from fetal bone marrow were flushed out with Advanced DMEM/F-12, supplemented with 1% penicillin/streptomycin, 1% GlutaMAX, and HEPES 10 mM. Single liver cells or mononuclear cells from bone marrow were stained with an antibody cocktail to sort HSPCs as described before[18]. Single HSPCs (CD34+,lineage-, index sort) were sorted with the sony SH800S into round-bottom 384-well plates (Supplementary Fig. S1 online). HSPCs were cultured in StemSpan SFEM medium supplemented with growth factors as described before for 3-4 weeks before collection of the cells[18].

DNA isolation
DNA from skin biopsies, clonal ISC organoids and primary intestinal biopsies was extracted using Genomic tip 20/G (Qiagen). DNA from clonal HSPCs cultures, bulk preleukemic blast cells and T-cells was extracted using Qiamp DNA Micro Kit (Qiagen).

DS-associated myeloid preleukemia samples
Viable frozen peripheral blood samples from 2 DS patients with DS-associated myeloid preleukemia were obtained from the biobank commission of the Princess Máxima Center for Pediatric Oncology. Mononuclear cells were stained with a cocktail of the following antibodies: CD3-BV650 (Biolegend, Clone UCHT1, 300467, 1:100), CD4-PerCP/Cy5.5 (Biolegend, Clone OKT4, 317427, 1:200), CD8-BV785 (Biolegend, Clone SK1, 344739, 1:100), CD19-BV421 (Biolegend Clone HCD14, 30224, 1:100) , CD14-AF700 (Biolegend, Clone HCD14, 325614, 1:100), CD56-BV711 (Biolegend, Clone HCD56, 318335, 1:50), CD34-APC (Biolegend, Clone 561, 343607, 1:50), CD38-PE (Biolegend, Clone HIT2, 303505, 1:50) , CD33-PE/Cy7 (Biolegend, Clone WM53, 303433, 1:100), CD117-PE-dazzle594 (Biolegend, Clone 104D2, 1:100), CD16-FITC (Biolegend, Clone 3G8, 302005, 1:100), CD20-FITC (Biolegend, Clone IVB201, 302303, 1:100). Bulk T-cells (CD3+/ CD4+ and CD3+/CD8+) and preleukemic blast cells were sorted with the Astrios-EQ. Preleukemic blast cells were sorted according to the diagnostics flow data. Cell pellets were used for DNA isolation. Public data was used for the other 4 myeloid preleukemia samples[11].

Collection of post-infant data
Vcfs from cord blood and D21 post-infant HSPCs were obtained from Osorio et al 2018[4]. Vcfs from D21 post-infant ISCs were obtained from Blokzijl et al 2016[5].

Whole genome sequencing and read alignment
DNA libraries for Illumina sequencing were generated using standard protocols (Illumina) from 20 - 50 ng of genomic DNA isolated from clonally expanded haematopoietic blood and progenitor cells, preleukemic blast cells and T-cells. DNA libraries for Illumina sequencing from skin biopsies and clonal ISC organoids were generated from 500ng DNA. All samples were sequenced (2 x150 bp) using Illumina HiSeq X Ten sequencers or Nova sequencers to 30x base coverage.

Version 2.6.0 of the Illumina Analysis Pipeline (https://github.com/UMCUGenetics/IAP) was used to align the reads and call variants similar to5. Copy numbers and b-allele frequencies were in concordance with the trisomy and sex state of all the bulks and clones. Initiation files are available upon request.

The bulk skin biopsies of N01, NR1 and NR2 were sequenced on both the Illumina HiSeq X Ten sequencers and the Nova sequencers. The resulting BAM files were merged using samtools merge[39]. The library (LB) and sample (SM) fields of the header were unified for each readgroup in the new bamfile.

Base substitution filtering
Unique base substitutions, not present in bulk tissue were filtered similarly as described before5. We considered variants that were passed by VariantFiltration and had a GATK phred-scaled quality score (QUAL) ≥50 and MQ≥60. Variants with multiple alternative alleles were removed. We excluded variant positions that overlapped with single-nucleotide polymorphisms (SNPs) in the SNP database (dbSNP)

v137.b37, unless that variant had a COSMIC id (v76)[40,41]. In addition, we removed all variants that overlapped with an inhouse blacklist (available upon request). We only retained autosomal and X-chromosome variants. We additionally filtered on genotype quality (GQ), Depth (DP) and VAF. For bulk tissues we filtered on GQ≥10 and VAF=0, while for clones we used GQ≥99 and VAF>0.1. In both the bulk tissue and clone we used DP≥20. Bulk skin was used to for all fetuses to control for germline variants, except for fetus MH3 and MH2, for these fetuses we used bulk intestine.

We used Dirichlet modeling to check the clonality of the clones. Subsequently, we removed all variants with a VAF below 0.3 to retain only the clonal substitutions. For T21 samples, we used a VAF ≥ 0.2 on chromosome 21, to account for the different expected VAF of clonal mutations. For X chromosomal variants in male donors we used VAF≥0.99 and GQ≥10 for clones and DP≥10 for both clones and bulk tissues.

To identify variants that were (sub)clonally present in the bulk tissue, we first applied our filters as described above, but did not yet filter on QUAL, GQ, DP or VAF to generate a “somatic” vcf file. All obtained variants were characterized in the clones. For each variant we divided the clones into ‘present’ and ‘absent’ based on their genotype. We filtered the ‘present’ clones using the GQ, DP and VAF filters, we previously used for the clones, while we used the bulk GQ, DP and VAF filters for the ‘absent’ clones. If at least one ‘present’ clone and one ‘absent’ clone passed the filtering, the variant was retained. This way variants are retained that are both confidently present and confidently absent in at least one clone. Finally, all variants were manually inspected using IGV (v2.4.15)[42].

Indel calling / filtering

Indels were filtered similarly to SNVs, except for the following differences. Variants lying within 100b of a called germline indel were removed. We filtered on QUAL≥250. For both bulk tissue and clones we filtered on GQ≥99.

Structural variant calling / filtering

We ran Gridss (v2.2.2) with bwa (v0.7.17) to detect structural variants (SVs)[43,44]. The output was filtered using a public pool of normal (3792v1) file from the Hartwig medical foundation (HMF) with the structuralvariantannotation (commit: d6173c3d9dd1fa314c91092b51920925b22268c6) R package and code modified from the HMF pipeline. In addition, we filtered for somatic SVs by only retaining variants in which at least one clone had a quality of 0. Next, we calculated VAFs and kept only breakpoints for which at least one clone had VAF≥0.3. Then, all breakpoints were removed for which the partner was not kept. Finally, all variants were inspected by eye in IGV (v2.4.15)[42]. In the end, no SVs were observed.

Driver mutations

The mutation load per clone could potentially be influenced by somatic driver mutations. We checked for the presence of driver mutations in the identified somatic mutations. A mutation was considered as possible driver if it met two requirements. First it needed to be annotated with “MODERATE” or “HIGH” effect by snpeff(v4.1) and second it needed to either have a COSMIC id (v76) or be located in a gene that was annotated as somatic in the Cosmic cancer gene census (v88)[41,45].
The mutation load per clone could also be influenced by germline drivers. To identify potential predisposition variants, we started with the “somatic” vcf described before. We filtered the bulk tissue on GQ≥50, DP≥10 and VAF≥0.3. Next, we removed all variants that had an allele count of more than 10 in either The ExAC (annotated via dbNSFPv2.9) or the GoNL (v5) database[46,47,48]. Furthermore, we only retained variants that were annotated with “MODERATE” or “HIGH” effect by snpeff(v4.1). Additionally, all variants were removed that did not overlap with a cancer-genes list from Zhang et al. 2015[48]. After manual inspection, none of the remaining variants were determined to be driver mutations.

Mutation load accumulation

For each clone the total number of somatic mutations was extrapolated to the entire called genome, based on the surveyed fraction of the genome similar to[5]. For the comparison against post-infant data, we used only autosomal variants that were not present in corresponding bulk tissue in order to equally compare the mutations load with the same method. We calculated the lifelong mutation rate by dividing the mutation load with the age of the donor since conception. Next, a linear mixed-effects regression model was fitted to compare the mutation rates between the fetal, cord-blood and post-infant clones. A random intercept was modeled for the ‘donor’ to resolve the non-independence that results from having multiple measurements per donor. Significance values and 95% CI intervals were calculated using a two-tailed t-test.

To compare the mutation load between D21 and T21 fetal cells, we fitted a linear mixed-effects model where the extrapolated mutational load was fitted against the age of the donors, the trisomy state and the cell type. The trisomy state and the cell type were crossed. We allowed a different variance for D21 and T21 fetal cells. A random slope was modeled for the ‘donor’. We did not observe a difference in mutation load between the cell types or an interaction between cell type and trisomy state (Supplementary Fig. S3 online).

To test that the significance of our model was not dependent on its complexity, we fitted simplified versions of the model to our data. These also showed significant differences in mutation load between D21 and T21 (Supplementary Table S4 online). However, our original model had the best performance based on log likelihood, the Bayesian information criterion (BIC) and the Akaike information criterion (AIC).
To test whether the variance between D21 and T21 was different, a log-likelihood ratio test was used to compare our main model to a version with a single variance for both D21 and T21.

To test that the significance of our model was not dependent on a few single points, we performed a leave-n-out analysis. We iteratively removed each combination of n points from our data and fitted our model on the remaining data. We determined the distribution of *p*-values we got for each of our fixed variables. The difference between D21 and T21 fetal cells always remained significant for both n=1 and n=2 (Supplementary Fig. S4 online).

Outliers in the models were detected by calculating the odds of the standardized absolute residuals occurring under a standard normal distribution. Fdr values were calculated to correct for multiple testing.

To compare the number of early mutations between D21 and T21 fetal cells, we used mutations that were (sub-)clonally present in the bulk. As these mutations occurred before gastrulation, they should not be affected by donor age. Therefore, we fitted the same model as described previously on the mutation load of early mutations, but without age as an explanatory variable.

The Wilcoxon rank sum test with continuity correction was used to compared the number of dbs between D21 and T21 fetal cells. Mutations present in multiple clones of a single fetus were only counted in a single cell.

Construction of developmental lineage tree
We created a binary mutation matrix with size CxM, where M is the number of mutations and C the number of clones. One and zero indicate presence and absence of a mutation in a clone. A row with only zeros was added to the matrix to root the tree. The pairwise distances between the clones where then calculated and a neighbourhood tree was generated.

Next, we calculated the VAFs in the bulk samples for all identified somatic mutations. For each developmental lineage branch, we took the mutations with a non-zero VAF

and created a matrix containing the reference and alternative allele counts. Next, we performed a chi-square test on this matrix to see if the mutations had significantly different VAFs. This was the case for one branch, with three mutations, in which the mutations likely occurred in different cell divisions. After, we calculated the median VAF of the non-zero VAF mutations in each branch. We multiplied the median VAFs with factor 2, to get the contribution of these mutations to the bulk tissue.

To determine if the first two developmental lineage branches had different contribution to the bulk tissue, we calculated whether their VAFs, were significantly different. We summed up the reference and alternative allele counts of the non-zero VAF mutations in a branch. For each pairwise combination of branches we then performed a Fisher's exact test.

Mutational profile and signature analysis
For the comparisons against post-infant data, we used only autosomal variants, because X-chromosome variants were not called in the post-infant data. Furthermore, for the comparisons of D21 with T21 and the comparison of T21 with DS-associated myeloid preleukemia, we only used unique substitutions not present in the bulk tissue, because those mutations are most likely to have originated in the cell type of interest. Additionally, mutations subclonally present in bulk T-cells were not called in the DS-associated myeloid preleukemia samples. In other comparisons, all mutations were used. Because of the low mutational load per sample, mutations were pooled per category. Mutations occurring in multiple clones, were only counted once. Each mutational signature analysis, was performed by comparing two categories at a time. The T21 fetal HSPC with a high mutational load was not included in the T21 category, but was instead analyzed separately because it was an outlier. Chi-square tests were used to compare base substitution profiles. The mutational profiles were fitted to a matrix containing the COSMIC signatures and the recently discovered HSPC signature using Mutational Patterns[23]. To reduce overfitting, we applied an iterative reverse selection process. During each iteration the mutational profiles are fitted against the signatures. Next, the cosine similarity between the original and the reconstructed profile was calculated. The signature with the lowest contribution across the samples was removed. This process was repeated until the difference in cosine similarities between two iterations became more than the cutoff of 0.05.

To provide us with a confidence level of signature contributions, we performed a bootstrapped version of our signature refitting method, with 1000 iterations. For the bootstrapping we resampled the mutational profiles with replacement. By correlating the bootstrapped contribution of signatures, we were able to visualize how the selection of one signature influences the selection of another signature. SBS5 and SBS40, which have a cosine similarity of 0.83, are negatively correlated (Supplementary Fig. S8 online). This shows that a small difference in a mutational profile, can cause the signature refitting process, to select a different signature. Bootstrapping is thus necessary, to determine how confident the signature exposures are.

We used a permutation test to compare the mutational signatures between samples. This was done by permuting the mutation matrix 2000 times, while keeping the margins fixed. Refitting, was then performed on the permuted matrixes, using the signatures that were selected at least 50% of the time in the bootstrapping. This results in a distribution of exposures for each used signature. Next, we calculated per category and per signature, how often the permuted exposures were more extreme then the exposures calculated using the original matrix. This value was then divided by the number of permutations and multiplied by two, to generate a two-tailed *p*-value.

The previously described method to compare the signatures between two groups is rather stringent and removes signatures with a small contribution. As a result, SBS1 and SBS5 could no longer be detected in post-infant HSPCs, even though they were present in fetal HSPCs. Since the mutations caused by SBS1 and SBS5 can't disappear, we decided to use a less stringent refitting method. Since SBS1, SBS5 and the HSPC signature were found to be present in HSPCs, we refitted the D21 HSPCs, T21 HSPCs, DS-associated myeloid preleukemia blasts and the post-infant HSPCs with only these signatures using the standard method from MutationalPatterns.

Mutational spectra, using previously defined mutational contexts, for the dbs and indel mutations were generated using inhouse R scripts.

We used modified versions of functions from the MutationalPatterns package to test whether there was an enrichment of mutations in regulatory regions, exons or genes.

Data availability
Data are available on EGA under accession number EGAS00001003982. Additionally, Vcfs and mutation matrixes are provided via the Github repository described in 'Code availability'.

Code availability
Code can be found on github at
https://github.com/ToolsVanBox/Mutation_accumulation_T21.

Acknowledgements

We would like to thank the participants and the staff of clinic Gynaikon in Rotterdam for all the efforts and availability to collect and provide the fetal material; and M. Bialecka, M. Nieveen and V. Torrens Juaneda for technical help. We would like to thank P. Lijnzaad for our discussions on the linear mixed-effects models.

Author contributions

K.A.L.H., M.V. and S.M.C.S.L. performed sample isolation. K.A.L.H. and T.P. performed fluorescence-activated cell sorting (FACS). K.A.L.H., M.V. and E.K. performed clonal expansions and supervised sequencing. K.A.L.H., F.M. and R.B. wrote the manuscript. F.M. performed bioinformatic analyses. S.M.C.S.L. and M.L.H. collected fetal material. R.B. designed and supervised the study.

Additional Information

Competing interests
Authors declare no competing interests

Funding

This study was financially supported by a grant from the Dutch Cancer Society (KWF no. 11307) to R.B. and S.M.C.S.L. and a VIDI grant from the Netherlands Organisation for Scientific Research (NWO no. 016.Vidi.171.023) to R.B.

References

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. Nature 458, 719–724 (2009).
2. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science (80-.). 347, 78–81 (2015).
3. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science (80-.). 355, 1330–1334 (2017).
4. Osorio, F. G. et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. 25, 2308-2316.e4 (2018).
5. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature 538, 260–264 (2016).
6. Rozhok, A. I., Salstrom, J. L. & DeGregori, J. Stochastic modeling reveals an evolutionary mechanism underlying elevated rates of childhood leukemia. Proc. Natl. Acad. Sci. U. S. A. 113, 1050–1055 (2016).
7. Greaves, M. A causal mechanism for childhood acute lymphoblastic leukaemia. Nat. Rev. Cancer 18, 471–484 (2018).
8. Hasle, H., Clemmensen, I. H. & Mikkelsen, M. Risks of leukaemia and solid tumours in individuals with Down's syndrome. Lancet 355, 165–169 (2000).

9. Khan, I., Malinge, S. & Crispino, J. Myeloid leukemia in Down syndrome. Crit. Rev. Oncog. 16, 25–36 (2011).
10. Klusmann, J.-H. et al. Treatment and prognostic impact of transient leukemia in neonates with Down syndrome. Blood 111, 2991–2998 (2008).
11. Yoshida, K. et al. The landscape of somatic mutations in Down syndrome – related myeloid disorders. 45, (2013).
12. Labuhn, M. et al. Mechanisms of Progression of Myeloid Preleukemia to Transformed Myeloid Leukemia in Children with Down Syndrome. Cancer Cell 36, 123-138.e10 (2019).
13. Hitzler, J. K., Cheung, J., Li, Y., Scherer, S. W. & Zipursky, A. GATA1 mutations in transient leukemia and acute megakaryoblastic leukemia of Down syndrome. (2003). doi:10.1182/blood-2003-01-0013
14. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 574, 532–537 (2019).
15. Sheltzer, J. M. et al. Aneuploidy Drives Genomic Instability in Yeast.
16. Wang, Y. & Navin, N. E. Advances and Applications of Single-Cell Sequencing Technologies. Molecular Cell 58, 598–609 (2015).
17. Jager, M. et al. Measuring mutation accumulation in single human adult stem cells by whole-genome sequencing of organoid cultures. Nat. Protoc. 13, 59–78 (2017).
18. Huber, A. R., Manders, F., Oka, R. & van Boxtel, R. Characterizing Mutational Load and Clonal Composition of Human Blood. J. Vis. Exp. (2019). doi:10.3791/59846
19. Kuijk, E. et al. Early divergence of mutational processes in human fetal tissues. Sci. Adv. 5, eaaw1271 (2019).
20. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature 513, (2014).
21. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature 543, 714–718 (2017).
22. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. Nature (2018). doi:10.1038/s41586-018-0497-0
23. Alexandrov, L. B. et al. The Repertoire of Mutational Signatures in Human Cancer. bioRxiv 322859 (2018). doi:10.1101/322859
24. Maura, F. et al. A practical guide for mutational signature analysis in hematological malignancies. Nat. Commun. 10, 2969 (2019).
25. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. Nat. Genet. 47, 1402–1407 (2015).
26. Viel, A. et al. A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. EBioMedicine 20, 39–49 (2017).
27. Busciglio, J. & Yankner, B. A. Apoptosis and increased generation of reactive oxygen species in down's syndrome neurons in vitro. Nature 378, 776–779 (1995).
28. Roy, A. et al. Perturbation of fetal liver hematopoietic stem and progenitor cell development by trisomy 21. Proc. Natl. Acad. Sci. U. S. A. 109, 17579–84 (2012).
29. Banno, K. et al. Systematic Cellular Disease Models Reveal Synergistic Interaction of Trisomy 21 and GATA1 Mutations in Hematopoietic Abnormalities. Cell Rep. 15, 1228–1241 (2016).
30. Tunstall-Pedoe, O. et al. Abnormalities in the myeloid progenitor compartment in Down syndrome fetal liver precede acquisition of GATA1 mutations. Blood 112, 4507–4511 (2008).
31. Rossi, D. J., Jamieson, C. H. M. & Weissman, I. L. Leading Edge Review Stems Cells and the Pathways to Aging and Cancer. doi:10.1016/j.cell.2008.01.036
32. Cheshier, S. H., Morrison, S. J., Liao, X. & Weissman, I. L. In vivo proliferation and cell cycle kinetics of long-term self-renewing hematopoietic stem cells. Proc. Natl. Acad. Sci. 96, 3120–3125 (1999).
33. Taylor, A. M. et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell 33, 676-689.e3 (2018).
34. Morawiec, Z. et al. DNA damage and repair in children with Down's syndrome. Mutat. Res. 637, 118–123 (2008).

35. Cabelof, D. C. et al. MYELOID NEOPLASIA Mutational spectrum at GATA1 provides insights into mutagenesis and leukemogenesis in Down syndrome. 114, 2753–2763 (2009).
36. Roper, R. J. & Reeves, R. H. Understanding the Basis for Down Syndrome Phenotypes. doi:10.1371/journal.pgen.0020050
37. Beach, R. R. et al. Aneuploidy Causes Non-genetic Individuality. Cell 169, 229-242.e21 (2017).
38. Hardy, K., Handyside, A. H. & Winston, R. M. L. The human blastocyst: Cell number, death and allocation during late preimplantation development in vitro. Development 107, 597–604 (1989).
39. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
40. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311 (2001).
41. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 45, D777–D783 (2016).
42. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192 (2012).
43. Cameron, D. L. et al. GRIDSS : sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. 2050–2060 (2017). doi:10.1101/gr.222109.117.Freely
44. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26, 589–595 (2010).
45. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 6, 80–92 (2012).
46. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Hum. Mutat. 34, E2393–E2402 (2013).
47. Consortium, T. G. of the N. et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46, 818 (2014).
48. Zhang, J. et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N. Engl. J. Med. 373, 2336–2346 (2015).

Supplementary material

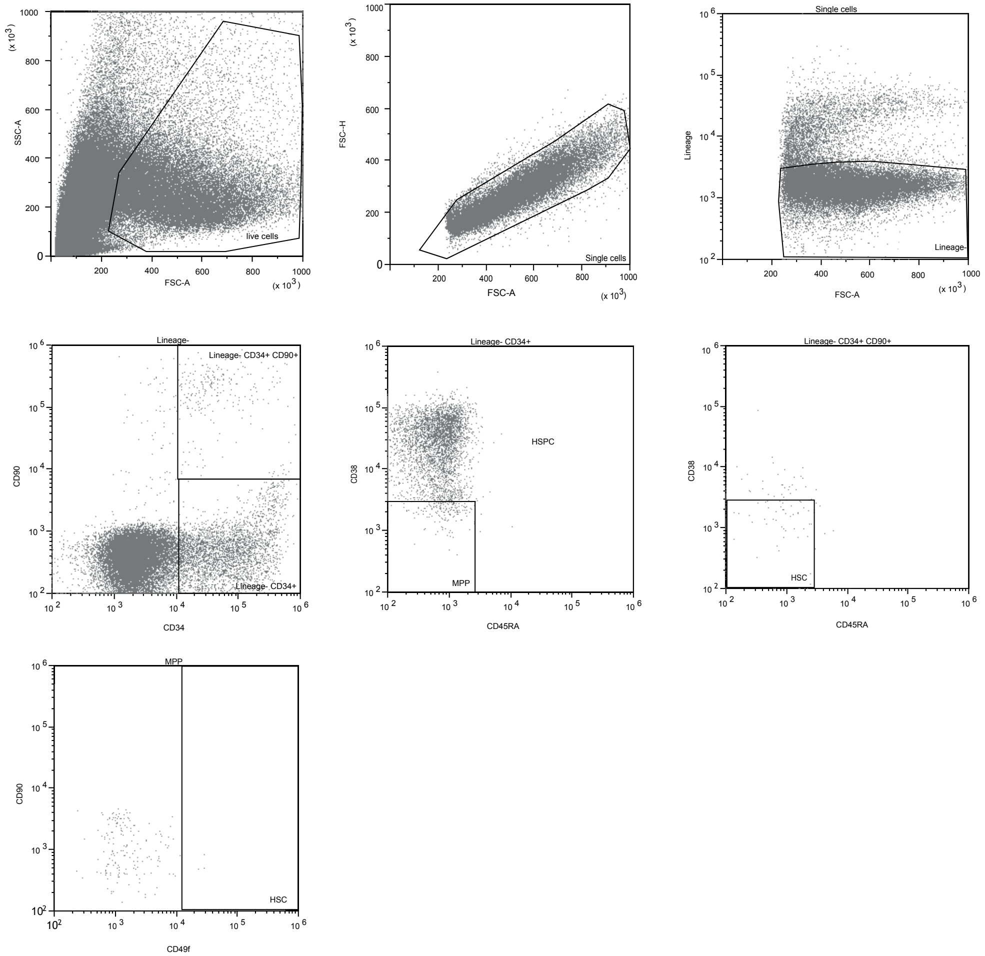
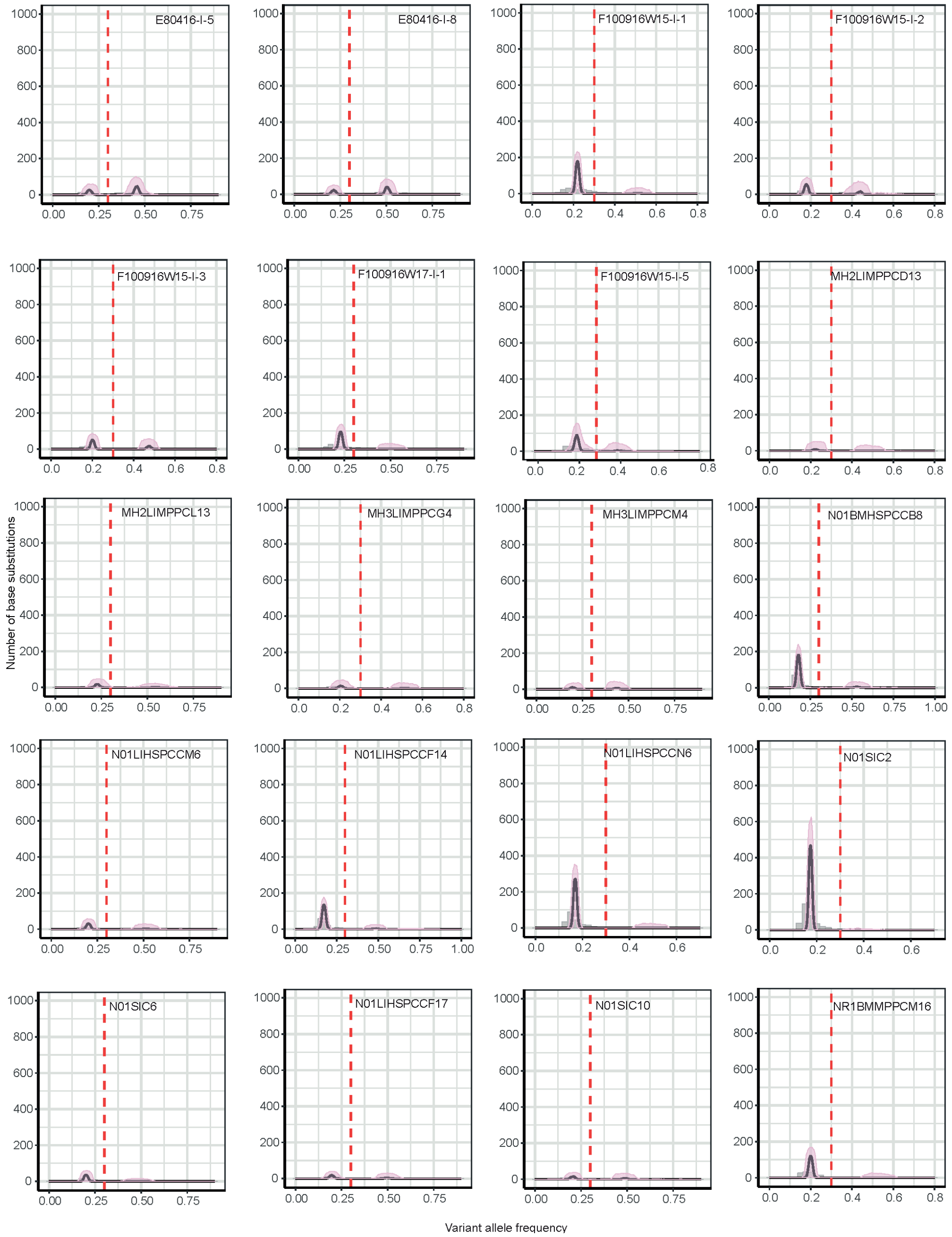
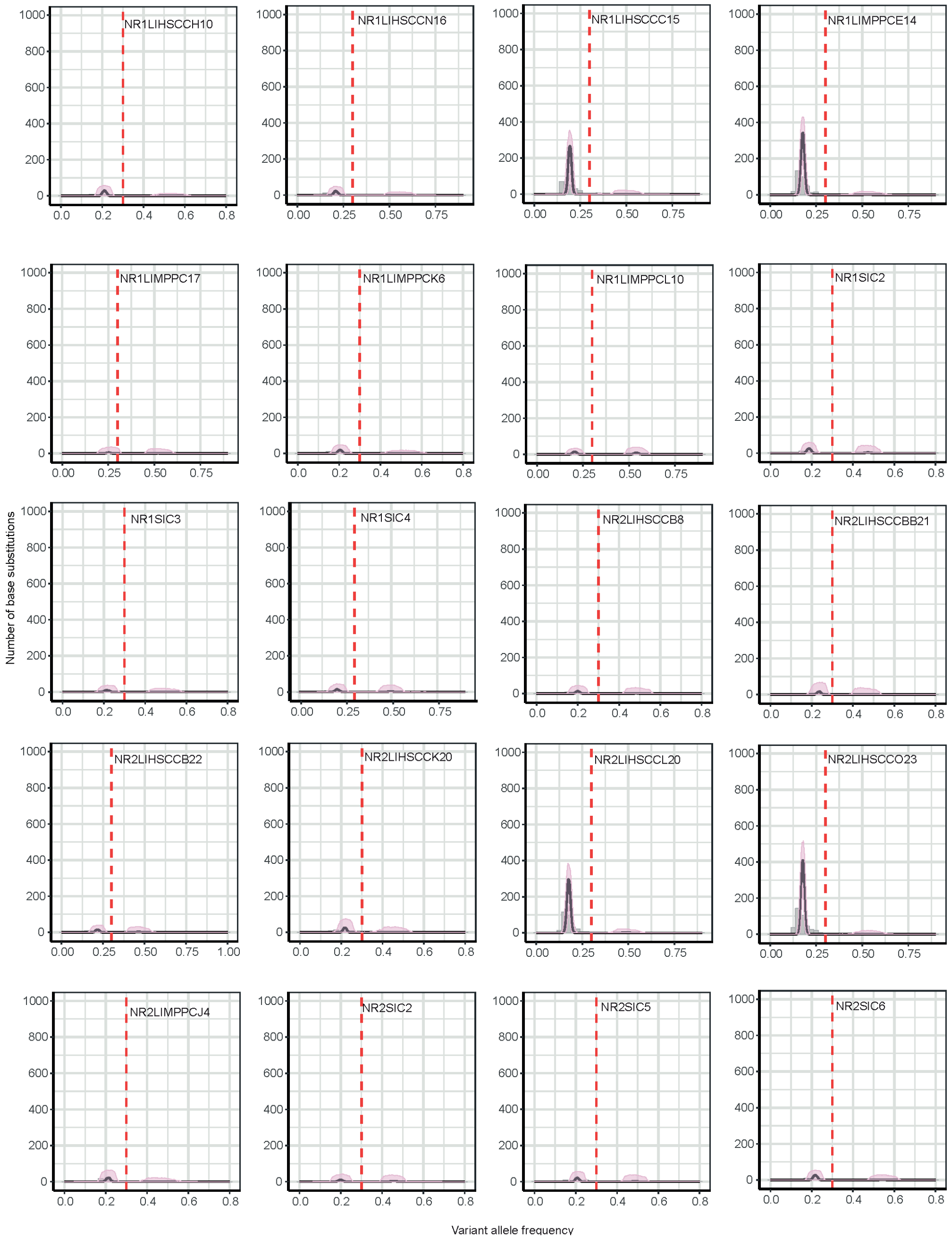


Fig. S1 Haematopoietic stem and progenitor cells isolation strategy.
Representative FACS strategy to sort haematopoietic stem and progenitor cells from fetal liver and bone marrow. Example data is from a trisomy 21 (T21) fetal liver


E80416-I-5
E80416-I-8
F100916W15-I-1
F100916W15-I-2
F100916W15-I-3
F100916W17-I-1
F100916W15-I-5
MH2LIMPPCD13
MH2LIMPPCL13
MH3LIMPPCG4
MH3LIMPPCM4
N01BMHSPCCB8
N01LIHSPCCM6
N01LIHSPCCF14
N01LIHSPCCN6
N01SIC2
N01SIC6
N01LIHSPCCF17
N01SIC10
NR1BMMPPCM16
Number of base substitutions
Variant allele frequency
0
200
400
600
800
1000
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8


NR1LIHSCCH10
NR1LIHSCCN16
NR1LIHSCCC15
NR1LIMPPCE14
NR1LIMPPC17
NR1LIMPPCK6
NR1LIMPPCL10
NR1SIC2
NR1SIC3
NR1SIC4
NR2LIHSCCB8
NR2LIHSCCBB21
NR2LIHSCCB22
NR2LIHSCCK20
NR2LIHSCCL20
NR2LIHSCCO23
NR2LIMPPCJ4
NR2SIC2
NR2SIC5
NR2SIC6
Number of base substitutions
Variant allele frequency
0
200
400
600
800
1000
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.6
0.8

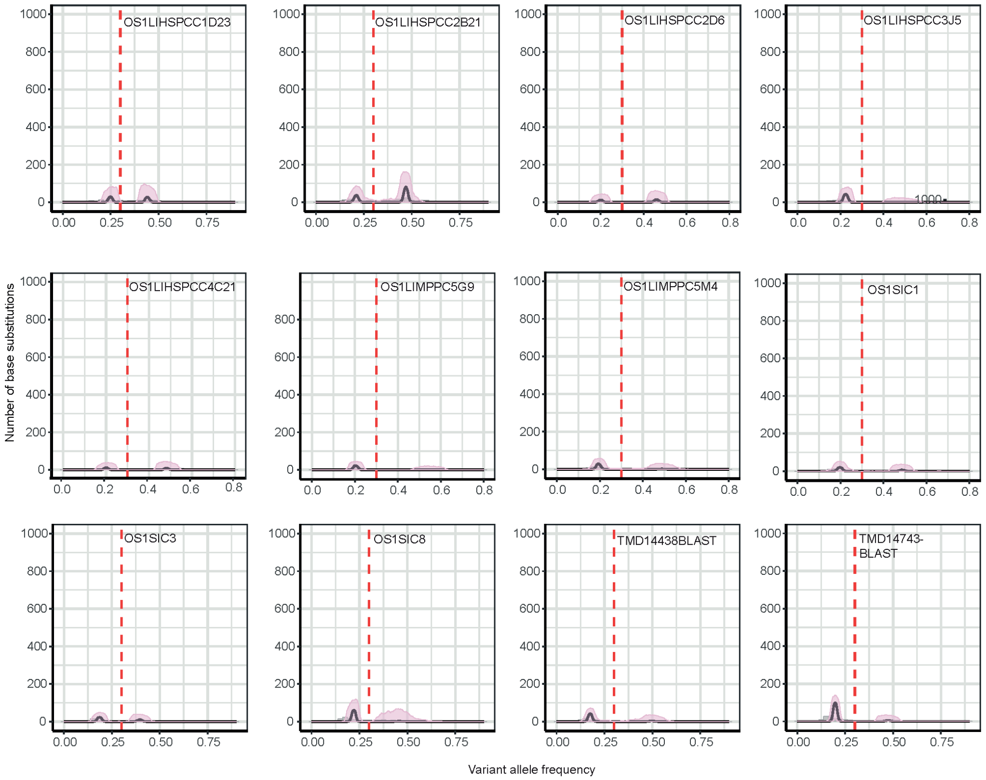


Fig. S2. Variant allele fractions (VAF) of base substitutions in sequenced clones.
Dirichlet modeling was used to determine the clonality of the cells. Histograms of the variant allele frequency of each sequenced sample to detect clonal single base substitutions. Clonal heterozygous mutations peak at VAF = 0.5. A threshold of VAF 0.3 was used to obtain mutations that were clonal and present in the original haematopoietic stem and progenitor cells or intestinal stem cells. Clonal mutations in the DS-associated myeloid preleukemia samples indicate the mutations present in the cell which underwent clonal expansion. Mutations acquired during or after clonal culture have lower VAFs and are therefore excluded. Shaded area represents the 95% posterior confidence intervals for the fitted distribution (pink area). In most samples, two clusters of mutations can be identified.

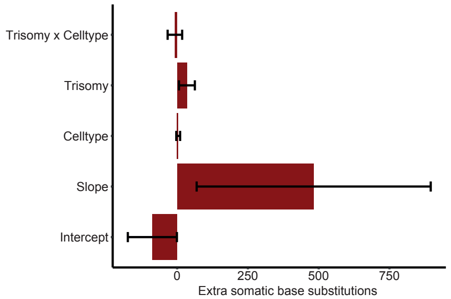


Fig. S3. Model parameters to determine somatic mutation load in fetal stem cells.
Model parameters of the linear mixed-effects model comparing the mutation load of disomy 21 (D21) vs trisomy 21 (T21). The model estimates of the explanatory variables are shown. Error bars represent 95% confidence intervals.

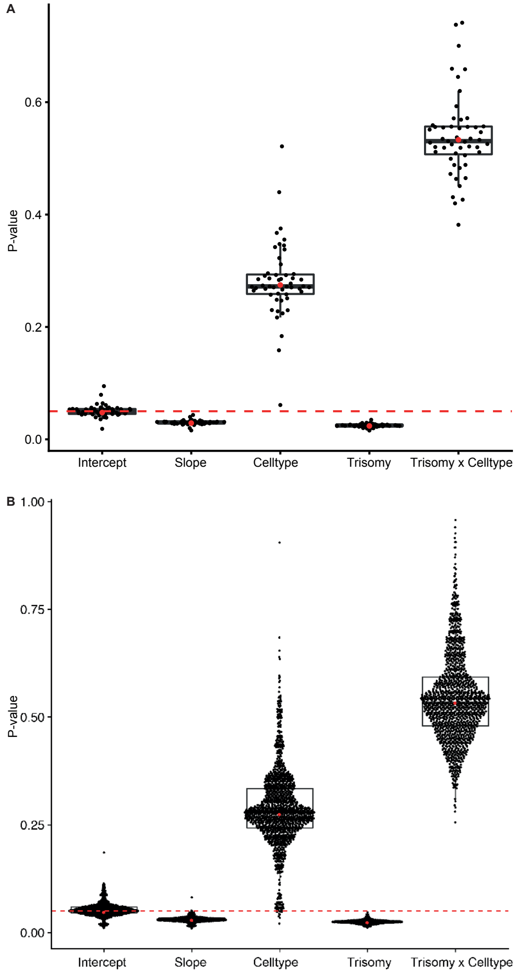


Fig. S4. Leave-n-out analysis on disomy 21 (D21) and trisomy 21 (T21) fetal stem and progenitor cells.
Each combination of n-points is iteratively removed and the linear mixed-effects model is calculated on the remaining data. The resulting P-values are shown for the different explanatory variables of the model. **a** n=1. **b** n=2.

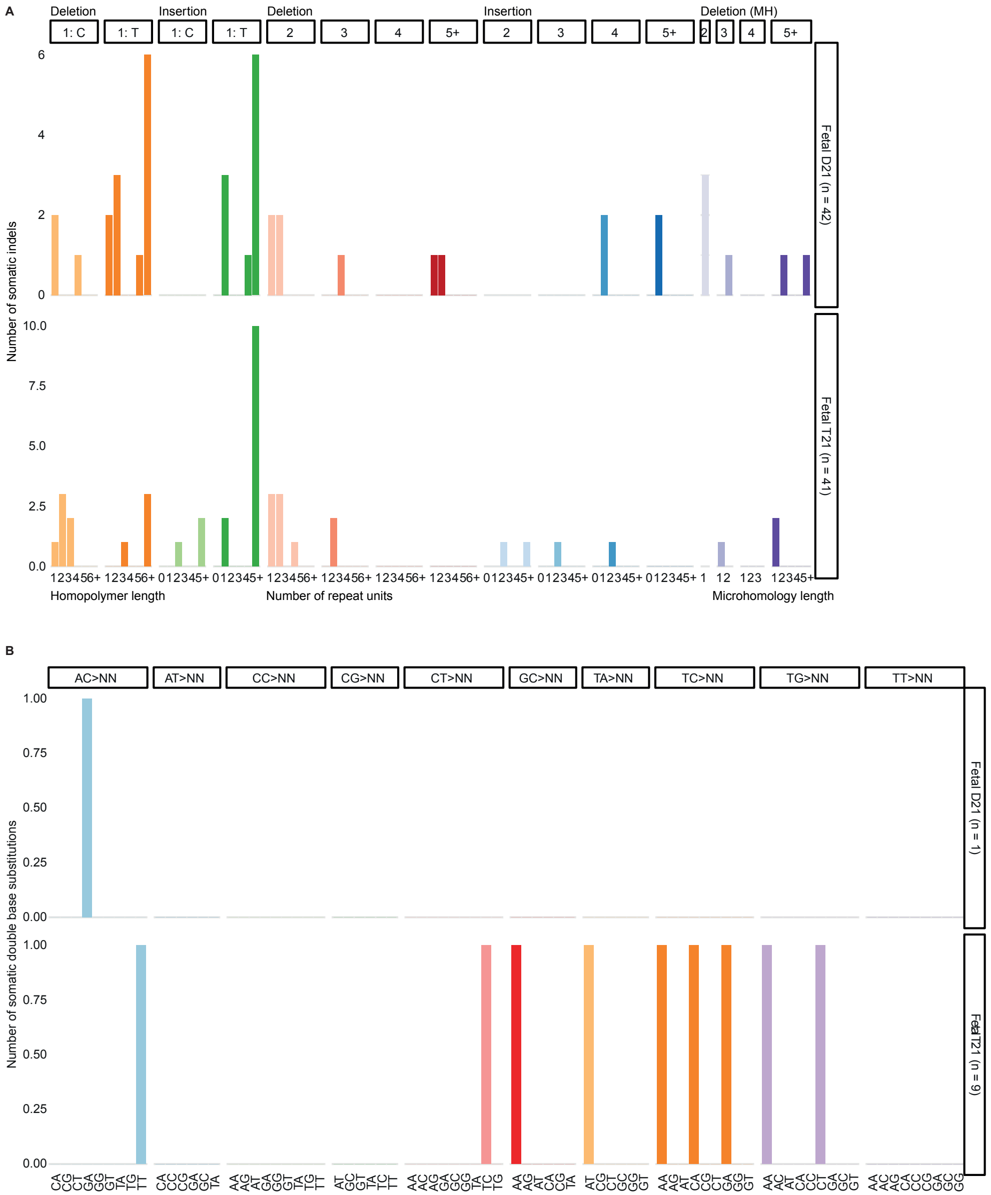


Fig. S5. Indel and double base substitution (dbs) spectra.
a Contribution of the indicated indel mutation types to the indel mutation spectrum in disomy 21 (D21) and trisomy 21 (T21) fetal stem and progenitor cells. The top of the plot shows the indel size. The bottom shows how often the mutated bases are repeated in the genome. Mutations are pooled per category (D21 fetal: 28 clones; 5 donors, T21 fetal: 23 clones; 4 donors). **b** Contribution of the indicated dbs mutation types to the dbs mutation spectrum in D21 and T21 fetal stem and progenitor cells. The top of the plot shows the reference bases. The bottom of the plot shows their substitution. Mutations are pooled per category (D21 fetal: 28 clones; 5 donors, T21 fetal: 23 clones; 4 donors).

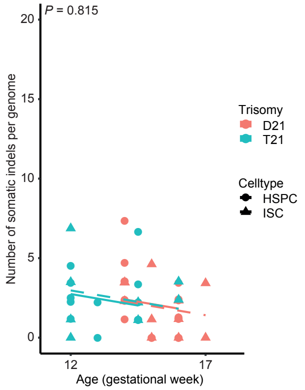


Fig. S6. Somatic indel mutation numbers in fetal stem and progenitor cells.
The number of somatic indels per genome plotted against the donor age (D21 fetal: 28 clones; 5 donors, T21 fetal: 23 clones; 4 donors). Dashed line: intestinal stem cells (ISC), full line: haematopoietic stem and progenitor cells (HSPC). P-value shows the difference between T21 and D21. (linear mixed-effects model, two-tailed t-test).

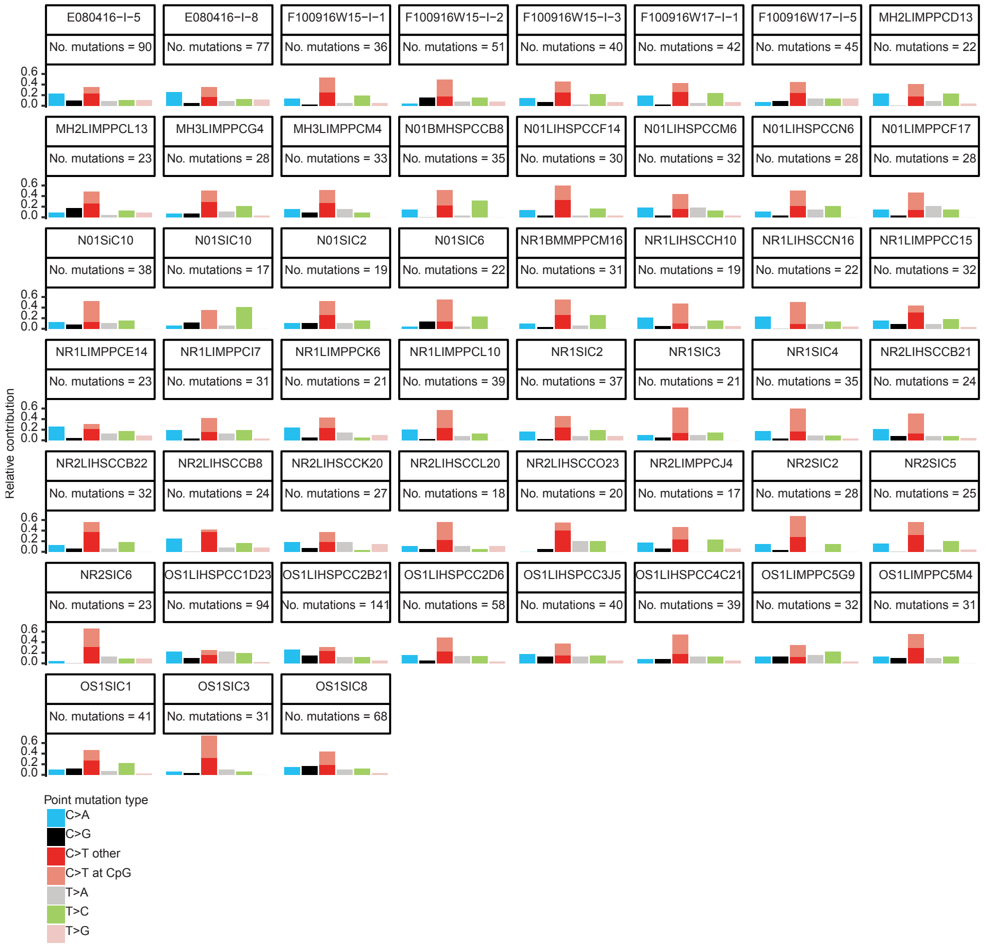


Fig. S7. 7-channel mutation spectra of each sequenced fetal clone.
Spectrum of point substitutions for each clone. The total number of point substitutions is indicated.

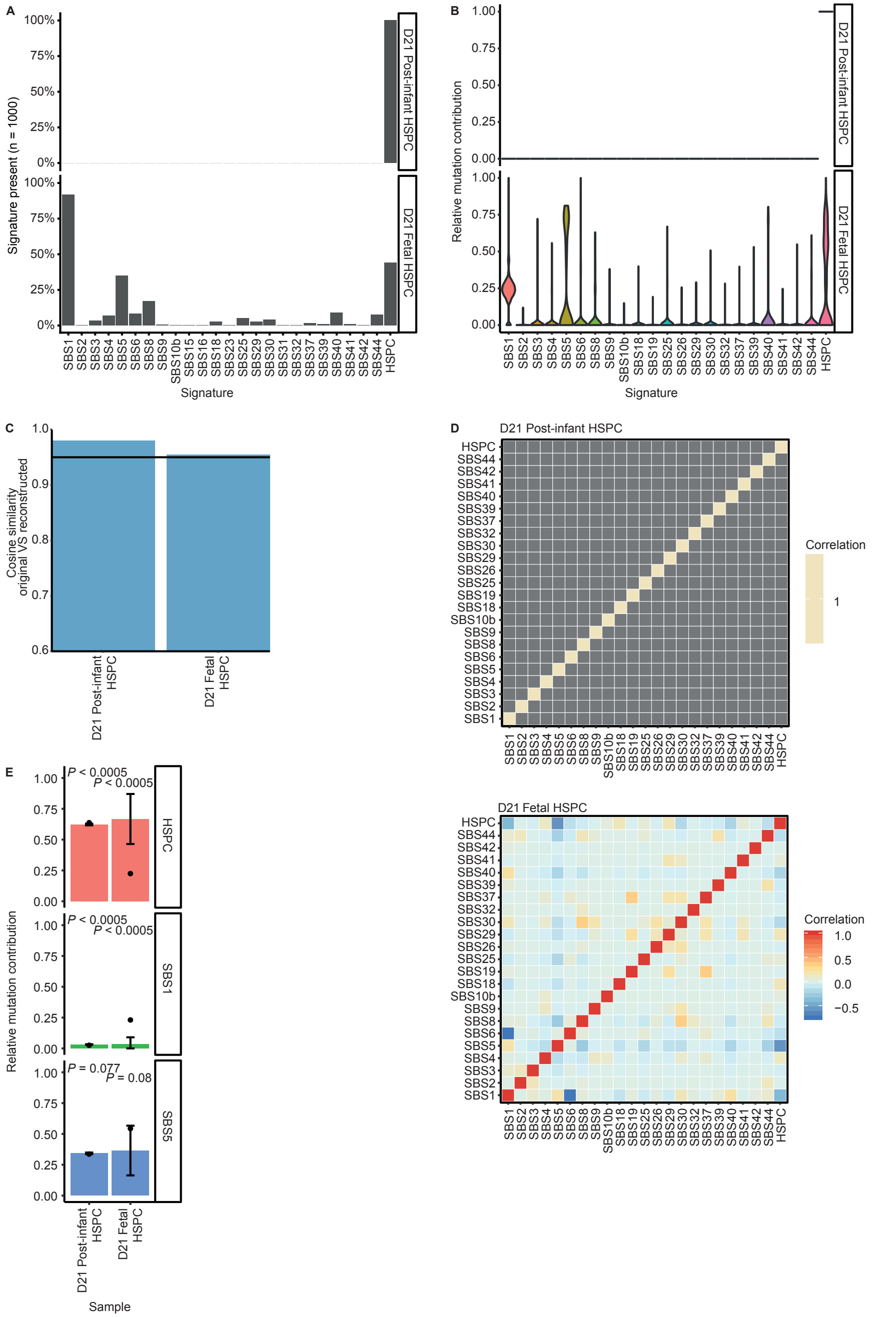


Fig. S8. Differences in mutational patterns between disomy 21 (D21) fetal haematopoietic stem and progenitor cells (HSPCs) and D21 post-infant HSPCs.
a For each signature the percentage of bootstrap iterations (1000 iterations) in which this signature was present is shown. **b** Violin plot of the bootstrapped (1000 iterations) number of base substitutions that each mutational signature contributed to the mutational profiles. Thicker parts of the violin are supported by more iterations of the bootstrap. The widths are scaled to the maximum density of each signature. **c** The cosine similarity between the mutational profiles and the mean reconstructed profiles, based on the

signature refitting are shown. **d** Heatmaps depicting the correlation of bootstrapped signature contributions are shown. **e** Signature permutation test (2000 permutations) for fetal vs post-infant D21 HSPCs. The bars show the mean relative signature contribution of the permutations. The error bars show the 2.5% and 97.5% quantiles. The dots show the actual measured relative signature contribution. (D21 Post-infant HSPC: n = 10924; 18 clones; 5 donors, D21 fetal HSPC: n = 353; 17 clones; 3 donors).

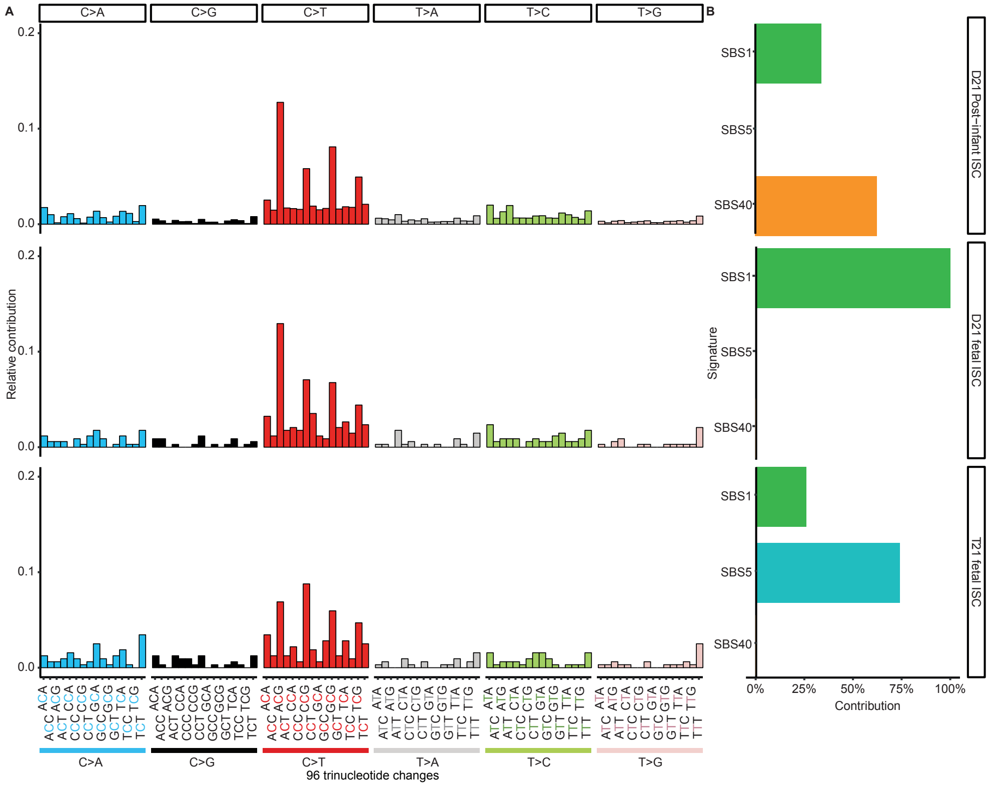


Fig. S9. Mutational patterns of disomy 21 (D21) post-infant, D21 fetal and trisomy 21 (T21) fetal intestinal stem cells (ISC).
a Spectra of point substitutions. The substitutions are pooled per category. (D21 Post-infant ISC: n = 21471; 14 clones; 9 donors, D21 fetal ISC: n = 340; 11 clones; 4 donors, T21 fetal ISC: n = 319; 9 clones; 3 donors). **b** The relative contribution of different mutational signatures to the spectra of point substitutions.

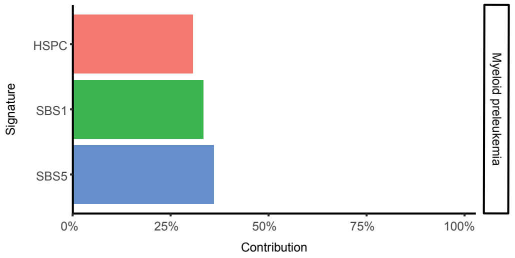


Fig. S10. Mutational signatures of preleukemic bulk blast cells from DS-associated myeloid preleukemia patients.
The relative contribution of different mutational signatures to the spectra of point substitutions. The substitutions from multiple samples were pooled together. (DS-associated myeloid preleukemia: n = 177; 6 donors).

For Table S1-S4 see: https://doi.org/10.1038/s41598-020-69822-1

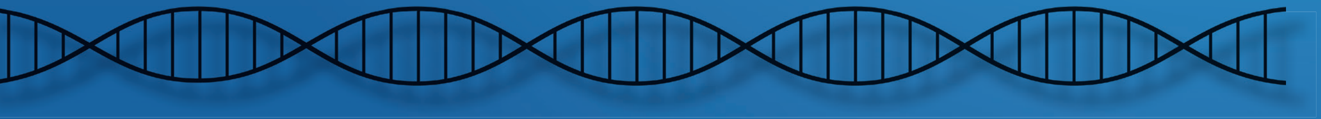
Chapter 3

MutationalPatterns: The one stop shop for the analysis of mutational processes

Freek Manders[1,2], Arianne M. Brandsma[1,2], Jurrian de Kanter[1,2], Mark Verheul[1,2], Rurika Oka[1,2], Markus J. van Roosmalen[1,2], Bastiaan van der Roest[2,3,4], Arne van Hoeck[2,3], Edwin Cuppen[2,3], and Ruben van Boxtel[1,2,*]

[1]Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584CS Utrecht, The Netherlands
[2]Oncode Institute, Jaarbeursplein 6, 3521 AL Utrecht, The Netherlands
[3]Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584, CG, Utrecht, The Netherlands
[4]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584, CG, Utrecht, The Netherlands
[*]Corresponding author: R.vanBoxtel@prinsesmaximacentrum.nl

BMC Genomics, 2022



Abstract

Background
The collective of somatic mutations in a genome represents a record of mutational processes that have been operative in a cell. These processes can be investigated by extracting relevant mutational patterns from sequencing data.

Results
Here, we present the next version of MutationalPatterns, an R/Bioconductor package, which allows in-depth mutational analysis of catalogues of single and double base substitutions as well as small insertions and deletions. Major features of the package include the possibility to perform regional mutation spectra analyses and the possibility to detect strand asymmetry phenomena, such as lesion segregation. On top of this, the package also contains functions to determine how likely it is that a signature can cause damaging mutations (i.e., mutations that affect protein function). This updated package supports stricter signature refitting on known signatures in order to prevent overfitting. Using simulated mutation matrices containing varied signature contributions, we showed that reliable refitting can be achieved even when only 50 mutations are present per signature. Additionally, we incorporated bootstrapped signature refitting to assess the robustness of the signature analyses. Finally, we applied the package on genome mutation data of cell lines in which we deleted specific DNA repair processes and on large cancer datasets, to show how the package can be used to generate novel biological insights.

Conclusions
This novel version of MutationalPatterns allows for more comprehensive analyses and visualization of mutational patterns in order to study the underlying processes. Ultimately, in-depth mutational analyses may contribute to improved biological insights in mechanisms of mutation accumulation as well as aid cancer diagnostics. MutationalPatterns is freely available at http://bioconductor.org/packages/MutationalPatterns.

Keywords
R, regional mutation patterns, mutagenic processes, mutational signatures, indels, base substitutions, somatic mutations

Background

Mutational landscapes in the genomes of cells are the result of a balance between mutagenic and DNA-repair processes[1]. The somatic mutations that shape these landscapes gradually accumulate throughout life in both healthy and malignant cells[2,3]. As a result, the complete collection of somatic mutations in the genome of a cell forms a record of the mutational processes that have been active throughout the life of that cell. In-depth analyses of somatic mutations can allow us to better understand the mutational processes that caused them[4].

First, such analyses can provide insight into the etiology of cancer by identifying mutagenic exposures, which ultimately contribute to the accumulation of cancer driving mutations. For example, we recently identified a mutational pattern caused by a carcinogenic strain of Escherichia coli found in the gut of ~20% of healthy individuals[5]. This pattern matched mutations found in colorectal cancer driver genes, indicating a direct role in tumorigenesis. Mutational patterns have been systematically determined *in vitro* for many environmental mutagenic agents, which can be used to deduce cancer causes[6]. The effects of such agents can also be found *in vivo*. For example, we recently found mutations caused by exposure to the antiviral drug ganciclovir, which patients received to treat a viral infection after a hematopoietic stem cell transplant[7]. Second, studying mutational processes can be useful for improved cancer diagnostics. For example, the presence of certain mutational signatures can be used as a functional readout for deficiency of homologous recombination (HR)-mediated double strand break repair[8,9]. Cancers with a defect in this repair pathway are selectively sensitive to poly(ADP-ribose) polymerase (PARP) inhibitors, providing a targeted therapy for the patients[10,11].

One of the most popular tools to analyze somatic mutation profiles is the R/Bioconductor package MutationalPatterns (v1.4.3), which can be used to easily investigate mutation spectra[12–19]. It can also be used to identify new signatures in mutation data using Nonnegative Matrix Factorization (NMF) and to determine the contribution of previously defined signatures to a sample using a method known as “signature refitting”[4]. However, the original version of this package has several limitations. First, the package is limited to single base substitutions (SBSs) and cannot be used for small insertions and deletions (indels) or double base substitutions (DBSs) even though signatures for these mutation types have recently been identified in large pan-cancer sequencing efforts[13]. The package also suffers from signature overfitting when determining the contribution of known patterns to a sample, which can result in too many signatures being attributed[20]. Additionally, the package only allows for analyzing spectra for mutations in the entire genome, making it difficult to study the involvement of specific genomic elements, such as enhancers or secondary hairpin structures. The ability to investigate the role of such elements in mutation accumulation is important, because this allows for identifying the molecular mechanisms by which certain processes induce mutagenesis[21–23].

Here we present a novel, almost completely rewritten version of MutationalPatterns (v3.4.0) for the analysis of mutational processes, which is easy-to-use and contains many new features, such as DNA lesion segregation[24]. Existing features have also been improved, resulting in a very comprehensive package that can be used for both basic and more advanced mutational pattern analyses. MutationalPatterns (v3.4.0) supports DBSs, multi base substitutions (MBSs) and indels, and can automatically extract all these mutation types from a single variant call format (VCF) file. The package can generate region specific spectra and signature contributions to study the varying activities of mutational processes across the genome. The package also generates more accurate results by supporting stricter signature refitting. This refitting can also be bootstrapped to determine the confidence of the results. Additionally, a process known as lesion segregation can be investigated.

The MutationalPatterns package (v3.4.0) can be used to generate novel biological insights, which we demonstrate by applying it to whole genome sequencing (WGS) data obtained from a lymphoblastoid cell line, in which specific DNA repair processes were deleted using CRISPR-Cas9 genome editing, as well as by applying the package on large cancer datasets. Additionally, we demonstrate that the package scales well on these large datasets. Finally, we show the improved accuracy of the stricter signature refitting using simulated data.

Implementation

Mutation profiles

MutationalPatterns uses mutations as its input data, which can be loaded into R from VCF files with the “read_vcfs_as_granges” function. MutationalPatterns (v3.4.0) supports SBSs, DBSs, MBSs and indels, whereas the original version only supported SBSs. Multiple mutation types are allowed to be present in a single VCF file so that users do not have to split them beforehand. A specific mutation type can be selected as an argument of the “read_vcfs_as_granges” function when reading in the VCF files. Alternatively, the “get_mut_type” function can be used on data that is already loaded in memory.

DBS and MBS variants can be called by various variant callers, such as the Genome Analysis ToolKit (GATK) Mutect2, in two different ways[25]. The variants can be called explicitly as DBS and MBS variants or as neighboring SBSs. A downside of the first approach is that neighboring germline and somatic mutations can be called as a single combined DBS or MBS, because the variants are compared to the reference instead of the control sample. MutationalPatterns (v3.4.0) supports both approaches. When the second approach is used, neighboring SBSs will be merged into somatic DBS or MBS variants.

Because they get merged, DBS and MBS variants are no longer incorrectly identified as separate SBSs by MutationalPatterns (v3.4.0). This improves the quality of the SBS profiles, as DBS and MBS mutations often have a very different context on account of them being caused by different processes[13] (Additional file 1: Figure S1).

The contexts of SBS, indel and DBS variants, as defined by the Catalogue of Somatic Mutations in Cancer (COSMIC) can be retrieved with fast vectorized functions, namely “mut_context”, “get_indel_context” and “get_dbs_context”. The context of SBS variants consisted of its direct 5’ and 3’ bases in the original package. These contexts were chosen because they are generally the most informative and adding more bases drastically increases the feature space, leading to sparsity[4]. Indeed, adding only one extra base to both the upstream and downstream context increases the number of features from 96 to 1536. However, with the increasing availability of large sequencing cohorts such large feature spaces have become more manageable, making it easier to examine nucleotide preference more upstream or downstream of the mutated base. Therefore, MutationalPatterns’ users can now choose any context size for SBSs. The mutation contexts can be used for custom analyses. Alternatively, the number of mutations per context can be counted, resulting in a count matrix, where each row is a context and each column a sample. These matrices are created with the “mut_matrix”, “mut_matrix_stranded”, “count_indel_contexts”, “count_dbs_contexts” and “count_mbs_contexts” functions. The “count_mbs_contexts” function uses the length of the MBSs, because to date no COSMIC consensus has been defined.

The count matrices can be plotted as spectra or profiles for all the mutation types (Fig. 1a, b, c). The SBS spectra can be displayed for the individual samples. Additionally, the error bars can be displayed as standard deviation, 95% confidence interval (CI) and the standard error of the mean. A count matrix with a larger context can be visualized using the new “plot_profile_heatmap” or “plot_river” functions (Fig. 1d, Additional file 1: Figure S2). This last function can be especially helpful to provide a quick overview of a mutation spectrum with a wider context. Mutation profiles can

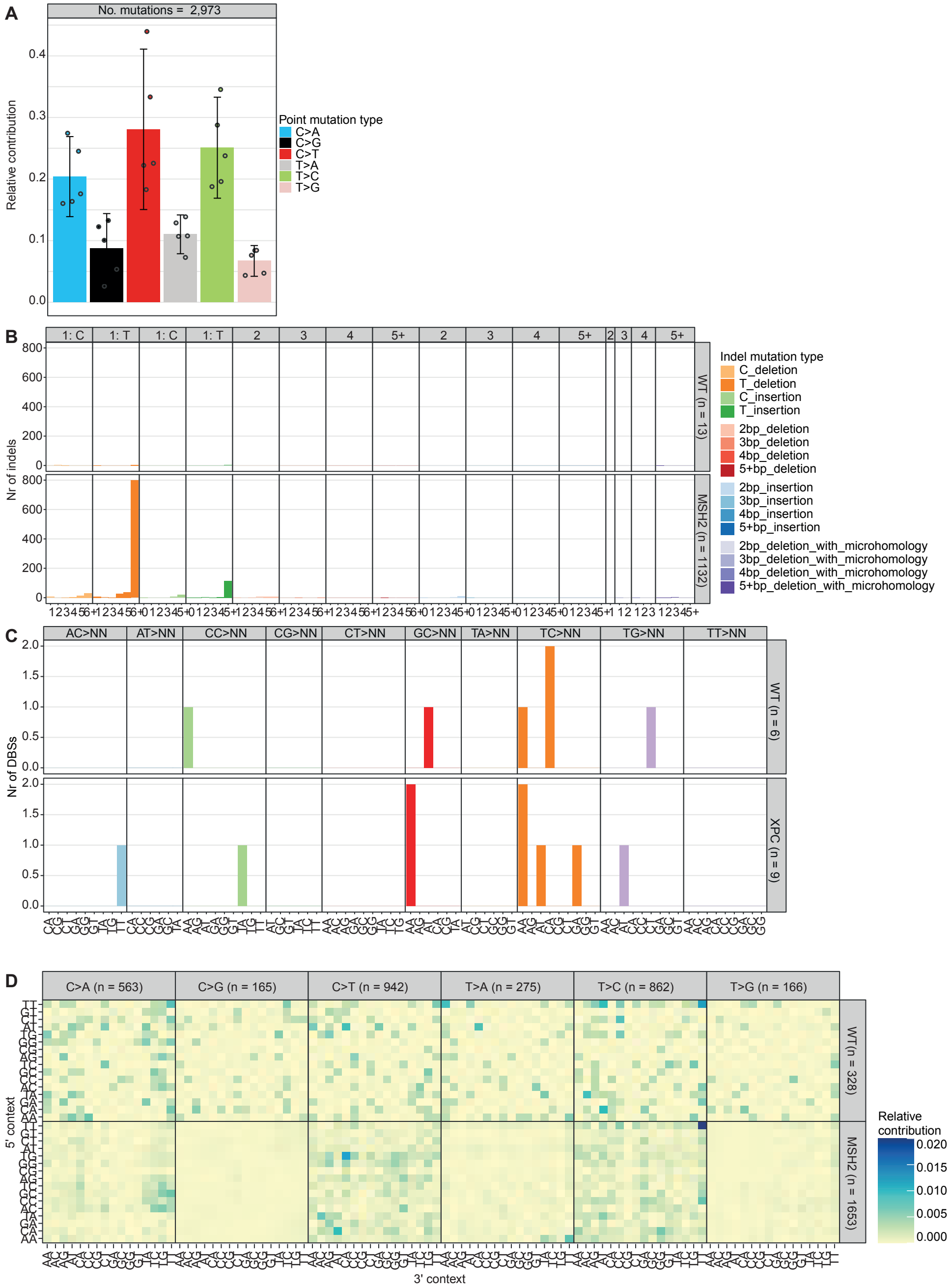


Fig. 1 Mutation profiles can be made for multiple mutation types.
a Relative contribution of the indicated mutation types to the point mutation spectrum. Bars depict the mean relative contribution of each mutation type over all the samples and error bars indicate the 95% confidence interval. The dots show the relative contributions of the individual samples. The total number of somatic point mutations per tissue is indicated. **b** Absolute contribution of the indicated mutation types to the indel spectrum for the wild-type (WT) and MSH2 knockout. The total number of indels per sample is indicated. **c** Absolute contribution of the indicated mutation types to the DBS spectrum for the wild-type (WT) and XPC knockout. The total number of DBSs per sample is indicated. **d** Heatmap depicting the relative contribution of the indicated mutation types and the surrounding bases to the point mutation spectrum for the WT and MSH2 knockout. The total number of somatic point mutations per tissue is indicated.

be compared using the "cos_sim_matrix" function, which calculates the cosine similarities between samples. The cosine similarity is a similarity score, that has a value between 0 and 1 and can be used to compare profiles with different amounts of mutations[4]. Next to visualizing or comparing them, a count matrix can also be used for downstream analyses, such as a de novo extraction of mutational signatures. In some cases, it can be useful to pool multiple samples within a count matrix to increase statistical power. This can be done using the new "pool_mut_mat" function.

Region specific analyses

Mutational processes can be influenced by regional genomic features at multiple scales, such as chromatin landscape, secondary hairpin structures as well as the major and minor groove of the DNA[21–23]. With the original version of MutationalPatterns (v1.4.3), it was possible to test for enrichment and/or depletion of the mutation load in such regions, using a Poisson test. However, the package lacked the possibility to automatically correct for multiple testing. In addition, mutational profiles in genomic regions could not be easily assessed. In MutationalPatterns (v3.4.0), multiple testing correction is now automatically performed by calculating the false discovery rate, when testing for enrichment and depletion[26]. In addition, multiple significance levels are now supported, which can be visualized using one or multiple asterisks. Furthermore, regional mutation profiles can be determined in detail. This is done by first splitting mutations based on pre-defined genomic regions, with the new "split_muts_region" function, which requires a GRanges or GRangesList object containing chromosome coordinates as its input. These coordinates can be read into R from file types like ".txt" or ".bed" files or they can be directly read from databases, such as Ensembl[27]. This analysis can be performed for multiple samples and multiple types of regions at once. A user could, for example, split a set of mutations into "promoter", "enhancer" and "other" mutations.

Splitting the mutations according to different genomic regions results in a GRangesList containing sample/region combinations. These combinations can be treated as separate samples by, for example, performing de novo signature analysis to identify processes that are specifically active in certain genomic regions. Knowing in which regions a signature is predominantly present, can lead to a better understanding of its etiology. Instead of treating the sample/region combinations as separate samples, the genomic regions can also be incorporated into the mutational contexts, using the new "lengthen_mut_matrix" function. This means that a mutational context like "A[C>A]A" could be split into “A[C>A]A-promoter” and “A[C>A]A-enhancer”. This analysis allows users to generate signatures that contain different mutation contexts in different genomic regions. Such signatures could be more specific than the regular COSMIC signatures.

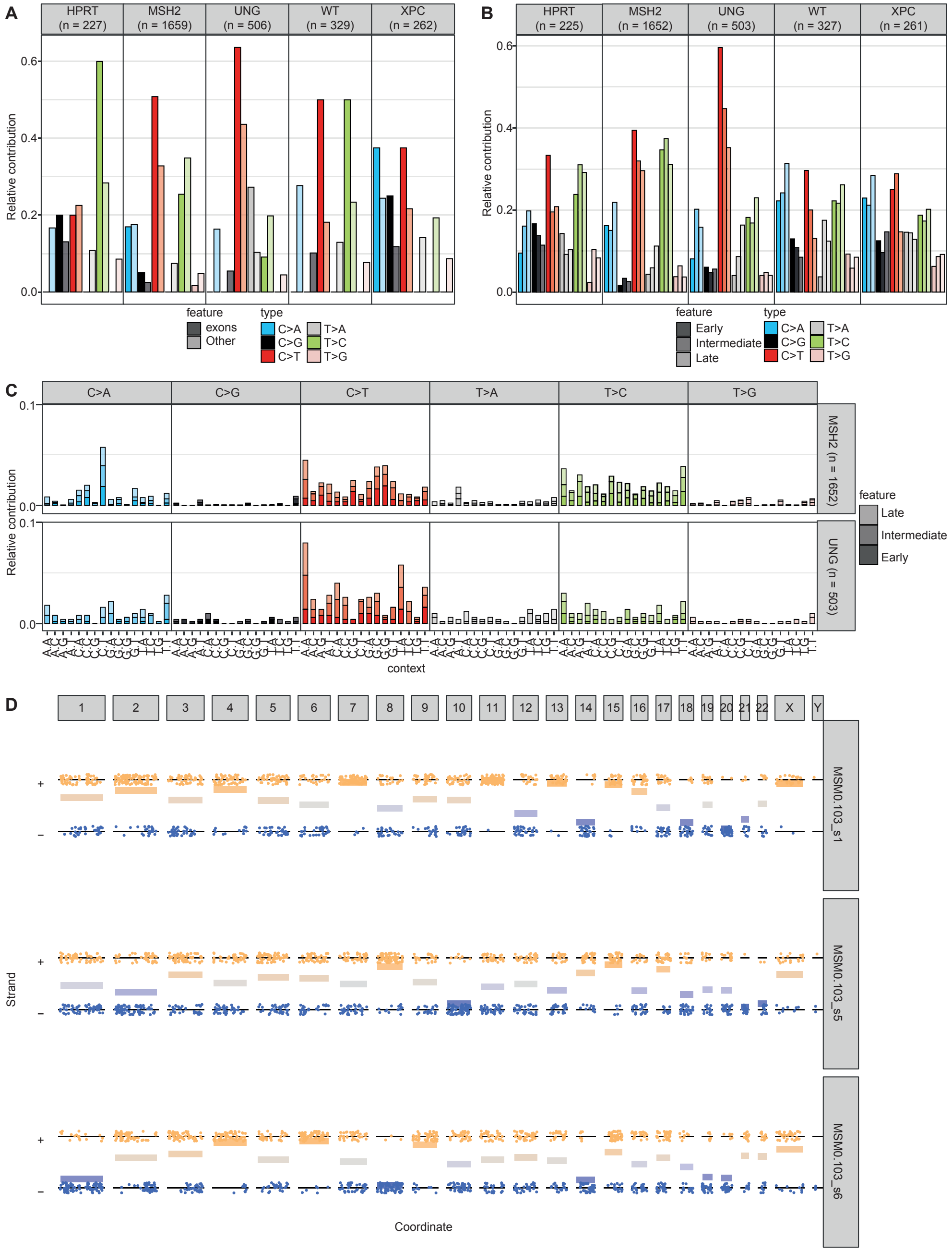


Fig. 2 Regional spectra show differences between genomic regions.
a Relative contribution of the indicated mutation types to the point mutation spectrum split between exons and the rest of the genome for each sample. The number of substitutions in each sample is indicated at the top of the figure. **b** Relative contribution of the indicated mutation types to the point mutation spectrum split between early-, intermediate-, and late-replicating DNA for each sample. The number of substitutions in each sample is indicated at the top of the figure. **c** Relative contribution of each trinucleotide change to the point mutation spectrum split between early- intermediate and late-replicating DNA for each sample. **d** A jitter plot depicting the presence of lesion segregation for each sample per chromosome. Each dot depicts a single base substitution. Any C>N or T>N is shown as a “+” strand mutation, while G>N and A>N mutations are shown on the “-“ strand. The x-axis shows the position of the mutations. The horizontal lines are calculated as the mean of the "+" and "-" strand, where "+" equals 1 and "-" equals 0. They indicate per chromosome on which strand most of the mutations are located. The mutations were downsampled to 33% to reduce the file size.

Region-specific mutation spectra can be visualized with the new “plot_spectrum_region” function, which contains the same arguments as the “plot_spectrum” function (Fig. 2a, b). In addition, region-specific 96-channel mutation profiles can be visualized with the new “plot_profile_region” function, which contains the same arguments as the “plot_96_profile” function (Fig. 2c). Both the “plot_spectrum_region” and “plot_profile_region” functions contain a “mode” argument, which allows users to normalize for the occurrence of the different mutation types per sample/region combination, per sample, or not at all.

Instead of using pre-determined genomic regions, it is also possible to compare the mutation spectra of regions with different mutation densities. These regions can be identified using the new “bin_mutation_density” function.

Regional mutational patterns can also be investigated using an unsupervised approach, which is unique to MutationalPatterns (v3.4.0), with the new “determine_regional_similarity” function. This function uses a sliding window approach to calculate the cosine similarity between the global mutation profile and the mutation profile of smaller genomic windows, allowing for the unbiased identification of regions with a mutation profile, that differs from the rest of the genome. Users can correct for the oligonucleotide frequency of the genomic windows using the “oligo_correction” argument. The function returns an S4 object, containing the genomic windows with their associated cosine similarities and the settings used to run the function. Because of the unbiased approach of this function, it works best on a large dataset containing at least 100,000 substitutions. The result of this analysis can be visualized using the new “plot_regional_similarity” function.

Lesion segregation

Mutation spectra sometimes contain Watson versus Crick strand asymmetries[24]. These asymmetries can be the result of many DNA lesions occurring during a single cell cycle. If these lesions are not properly repaired before the next genome duplication, then the resulting sister chromatids will segregate into different daughter cells, which will each inherit the lesions on opposite strands. This process is known as lesion segregation[24]. The presence of lesion segregation in mutation data can be calculated with the new "calculate_lesion_segregation" function. This calculation can be done for all mutations together or separately for the different mutation contexts. The results can be visualized using the "plot_lesion_segregation" function (Fig. 2d, Additional file 1: Figure S3).

Mutational signature analysis

When performing signature analyses, it is possible to either extract novel signatures using NMF, which is a type of dimensionality reduction[4], or to fit previously defined signatures to a mutation count matrix (signature refitting), using a non-negative least-squares optimization approach[28]. Both approaches could be applied on SBSs using the original MutationalPatterns (v1.4.3). With MutationalPatterns (v3.4.0), these approaches can be applied on all mutation types. By combining count matrices of different types, it is even possible to create a composite signature.

MutationalPatterns (v3.4.0) supports a variational Bayesian (Bayes) NMF algorithm from the ccfindR package to help choose the optimal number of signatures, in addition to the regular NMF algorithm[29] (Additional file 1: Figure S4). One challenge with de novo signature extraction is that extracted signatures can be very similar to previously defined signatures with known etiology. With the new "rename_nmf_signatures" function, these extracted signatures can be identified using cosine similarity scores and their names can be changed from an arbitrary naming to a custom naming that reflects their similarity to these previously defined signatures.

The original MutationalPatterns package already contained the "fit_to_signatures" function, which finds the optimal combination of signatures to reconstruct a profile and calculates a reconstructed profile based on this combination of signatures. However, this approach could lead to too many signatures being used to explain the data[20]. One simple method to reduce this overfitting, which was used in the vignette of the previous version of MutationalPatterns (v1.4.3), is to remove all signatures with less than 10 mutations. However, this method, which we will call "regular_10+", only reduced overfitting slightly. To reduce overfitting, we introduce the new "fit_to_signatures_strict" function. The default backwards selection method of this function iteratively refits a set of signatures to the data, each time removing the signature with the lowest contribution. During each iteration the cosine similarity between the original and reconstructed profile is calculated. The iteration process stops when the change in cosine similarity between two iterations is bigger than the user-specified

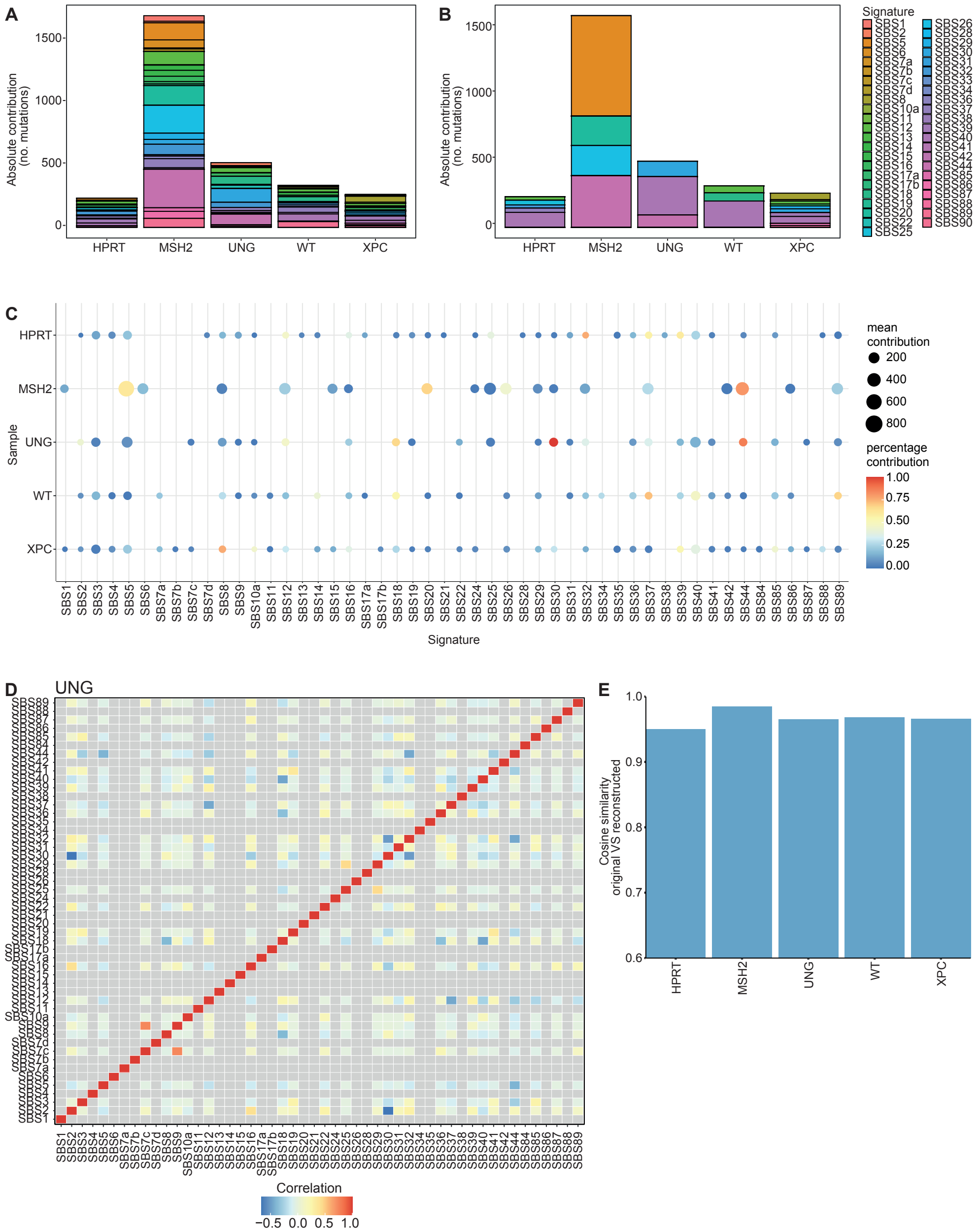


Fig. 3 Signature refitting is improved.

a Absolute contribution of each mutational signature for each sample using "regular" signature refitting and **b** "strict" signature refitting. **c** Dot plot showing the contribution of each mutational signature for each sample using bootstrapped signature refitting. The colour of a dot indicates the fraction of bootstrap iterations in which a signature contributed to a sample. The size indicates the mean number of contributing mutations across bootstrap iterations in which the contribution was not zero. **d** Heatmap depicting the Pearson correlation between signature contributions across the bootstrap iterations. **e** Bar graph depicting the cosine similarity between the original and reconstructed profiles of each sample based on signature refitting.

“max_delta” cutoff (Additional file 1: Figure S5). Users can set the “max_delta” cutoff based on their desired sensitivity and specificity. Stricter refitting, with this method, is comparable to a previously described approach and results in less signatures being chosen when tested on mutation data obtained from cell lines that lack specific DNA repair pathways (Fig. 3a, b; see Additional file 2)[13]. The “fit_to_signatures_strict” function also has a best subset selection approach. This method works similarly to the backwards selection approach. However, instead of removing the signature with the lowest contribution, each combination of x signatures is tried. This includes signatures that were not included in a previous iteration. Here, x is the number of signatures used during refitting, which is reduced by one in each iteration step. By default, “fit_to_signatures_strict” uses the backwards selection method, because the best subset method becomes very slow when fitting against more than 10-15 signatures. Therefore, we used the backwards selection method for all “strict” signature refitting analyses in the rest of this manuscript. Another way to reduce overfitting is to only use signatures that are known to be potentially active in your tissue/cells of interest. We recommend using this method in combination with “fit_to_signatures_strict” for optimal results.

In addition to estimating contributions of signatures to mutation spectra, it is also vital to know how confident these contributions are. The confidence of signature contributions can be determined using a bootstrapping approach with the new “fit_to_signatures_bootstrapped” function, which can use both the strict and the regular refitting methods. Its output can be visualized in multiple ways using the “plot_bootstrapped_contribution” function (Fig. 3c, Additional file 1: Figure S6). The signature contributions can be correlated between signatures across the different bootstrap iterations. This correlation can be visualized using the “plot_correlation_bootstrap” function (Fig. 3d). A negative correlation between two signatures means that each signature had a high contribution in iterations in which the other had a low contribution, which can occur when the refitting process has difficulty distinguishing between two similar signatures. One simple way to deal with highly similar signatures is to merge them. This can be done using the new “merge_signatures” function.

To test the accuracy of signature analysis, the cosine similarity between the reconstructed and original mutation profile needs to be determined. A high cosine similarity between the reconstructed and original profile indicates that the used signatures can explain the original spectrum well. This comparison between reconstructed and original mutation profiles can be visualized with the new “plot_original_vs_reconstructed" function (Fig. 3e).

In order to perform refitting, a matrix is required of the predefined signatures Signature matrices of COSMIC (v3.1 + v3.2), SIGNAL (v1) and SparseSignatures (v1) are now included in MutationalPatterns[6,13,15,30]. These matrices include general, tissue-specific and drug exposure signatures. The COSMIC matrices also include DBS and indel signatures, next to the standard SBS signatures. Signature matrices can be easily loaded using the new “get_known_signatures” function. By default, this function excludes several signatures from COSMIC and SIGNAL, because they are possible sequencing artefacts[13]. Users can choose to include these signatures, for example to check the quality of their data.

Signature-specific damaging potential analysis

Some signatures are more likely than others to have functional effects by causing premature stop codons (“stop gain”), splice site mutations or missense mutations, because of sequence specificity underlying these changes. With MutationalPatterns (v3.4.0) it is now possible to analyze how likely it is for a signature to either cause "stop gain", "missense", "synonymous" or "splice site" mutations for a set of genes of interest. For this analysis to be performed, the potential damage first needs to be calculated per mutational context, with the “context_potential_damage_analysis” function. Next, the potential damage per context is combined using a weighted sum to calculate the potential damage per signature using the “signature_potential_damage_analysis” function. The potential damage per signature is also normalized using a “hypothetical” flat signature, which contains the same weight for each mutation context.

This analysis will only take mutational contexts into account. Other features, such as open/closed chromatin, are not considered, because they vary per tissue type. However, this analysis can still give an indication of how damaging a signature might be, which could be supplemented by further custom analyses.

This new version of MutationalPatterns (v.3.4.0) also comes with many smaller updates and bugfixes. A comprehensive list can be found in Additional file 3: Table S1.

Results

Extended mutation context analysis and regional mutational patterns

To demonstrate the importance of analyzing extended mutation contexts, regional mutational patterns and lesion segregation for characterizing the underlying mutagenic processes, we applied MutationalPatterns (v3.4.0) to three published mutation datasets. First, we ran MutationalPatterns on 276 melanoma samples from

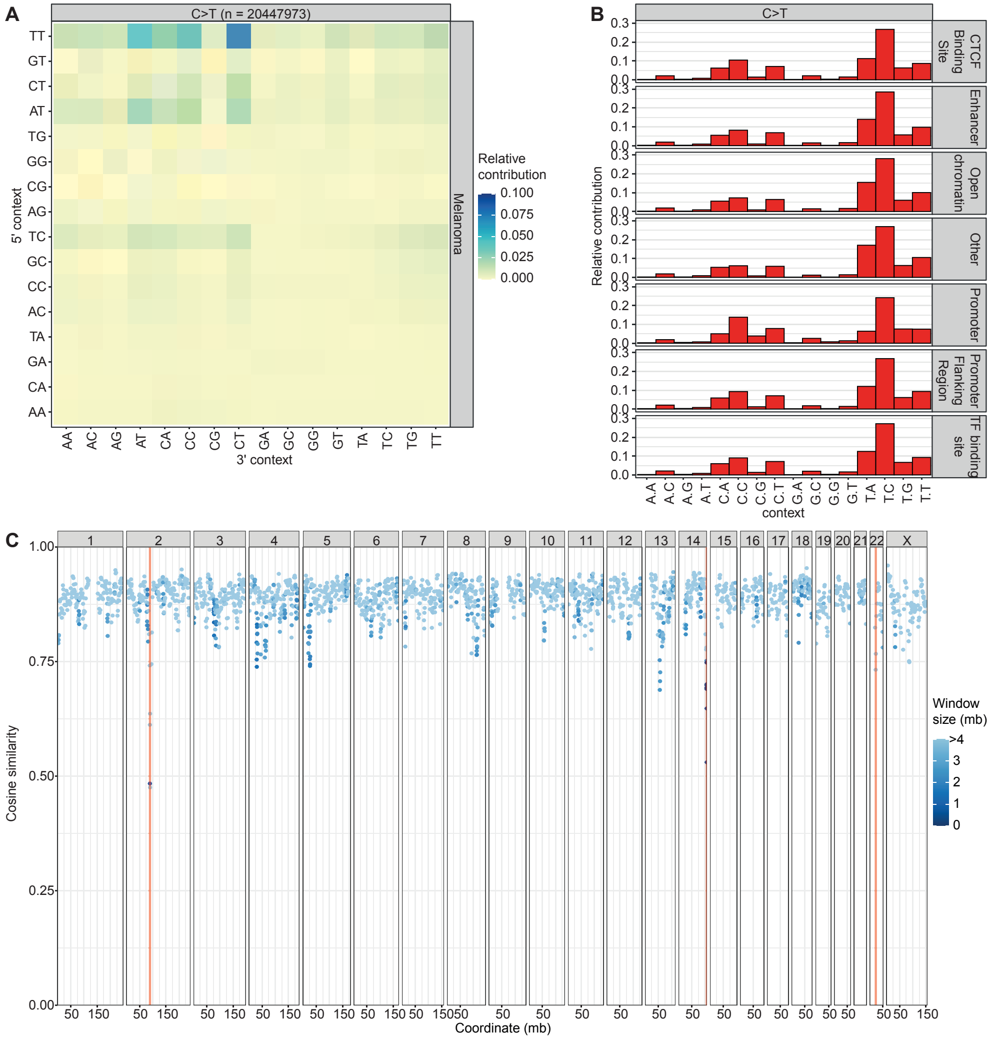


Fig. 4 Large cancer datasets show extended and regional mutation patterns.
a Heatmap depicting the relative contribution of the indicated mutation types and the surrounding bases to the point mutation spectrum for metastatic melanomas. The total number of somatic point mutations is indicated. Only C>T substitutions are shown, because other substitution types are much less common. **b** Relative contribution of each C>T trinucleotide change to the point mutation spectrum split between different genomic regions. **c** Graph depicting the similarity in the mutation profile between genomic windows and the rest of the genome. Each dot shows the cosine similarity between the mutation profiles of a single window and the rest of the genome. The dots are colored based on the sizes in mega bases of the windows. The IGK (chr2), IGH (chr14) and IGL (chr22) loci are visualized with vertical orange lines (46). The width of the lines is set at 1pt, because using the actual widths of these loci results in lines that are too small to be visible.

the Hartwig Medical Foundation (HMF) database. After pooling these samples, we observed that TT[C>T]CT mutations are the most common type of substitution (Fig. 4a, Additional file 1: Figure S7). This substitution type is more common than other

T[C>T]C substitutions, showing that the extended context has a large effect. Next, we compared the mutation patterns of the melanoma samples between the different genomic regions classified by the Ensembl regulatory build[31]. Interestingly, these patterns are very similar, suggesting that the epigenetic state of melanoma samples does not have a large effect on the types of mutations that occur in them (Fig. 4b).

Next, to show how MutationalPatterns (v3.4.0) can be used to identify regional activity of specific mutation processes in an unsupervised manner, we applied the package on 217 pooled pediatric B-cell Acute lymphoblastic leukemia (B-ALL) WGS samples[32]. These B-cell-derived leukemias have undergone VDJ recombination, which is associated with somatic hypermutation at loci encoding for immunoglobulin[33,34]. As somatic hypermutation is associated with a specific signature, these sites were expected to have a mutation spectrum that is different from the rest of the genome. Indeed, MutationalPatterns (v3.4.0) was able to detect a different spectrum for the two VDJ regions, located on chromosomes 2 and 14 (Fig. 4c). Some other regions also seem to have a different mutational pattern, several of which contain PCDH genes. However, further research is needed to explain these results. This example shows how MutationalPatterns (v3.4.0) can identify region-specific mutational processes in an unsupervised manner.

Finally, to show how MutationalPatterns (v3.4.0) can identify lesion segregation, we applied it on a dataset known to contain this phenomenon. We found significant lesion segregation (fdr = 4.41*10-116, 3.51*10-59 and 8.83*10-99, respectively) in data obtained from 3 samples of induced pluripotent stem cells treated with 0.109 uM of dibenz[a,h] anthracene diol-epoxide[6,24], using the "plot_lesion_segregation" function of MutationalPatterns (Fig. 2d). The rl20, which is a measure of lesion segregation, of these three samples was 22, 10 and 23, respectively. A value larger than 5 indicates the presence of lesion segregation[24]. It was even possible to spot sister-chromatid-exchange events, such as on chromosome 2 of sample MSM0.103_s6 (Fig. 2d, lower panel). To reduce the file size of the figure, 66% of the mutations of each sample were removed using the "downsample" argument of this function. Using MutationalPatterns (v3.4.0), we also found lesion segregation in patients that received the antiviral drug ganciclovir[7].

MutationalPatterns offers more functionality than other mutation analysis tools

An overview of the functions of MutationalPatterns (v3.4.0) and related tools is shown in Table 1. The original version of MutationalPatterns (v1.4.3) is also included in this table. An important advantage of the original package was that it combined many mutational analyses into a single package. This new version improves many of these features and adds many new and unique features.

Table 1: Feature comparison with other packages

Group	Feature	MutationalPatterns v3.4.0	MutationalPatterns v1.4.3 (12)	Sigprofiler (13)	Signature-Analyzer (13)	deconstructSigs (14)	sparseSignatures (15)	signeR (16)	somaticSignatures (17)	Maftools (18)	decompTumor2Sig (19)
Language	Language/platform	R (bioconductor)	R (bioconductor)	Python (+ R wrapper)	Python	R (cran)	R (bioconductor)	R (bioconductor)	R (bioconductor)	R (bioconductor)	R (bioconductor)
Genome	Supported genomes	Genome agnostic	Genome agnostic	Human, Mice, Rat, Yeast	-	Human	Genome agnostic	Genome agnostic	Genome agnostic	Genome agnostic	Genome agnostic
Mutation profile	96 SNV profile	X	X	X	-	X	-	X	X	X	X
	extended SNV profile	X	-	X	-	-	-	-	X	-	X
	Indel profile	X	-	X	-	-	-	-	-	-	-
	DBS profile	X	-	X	-	-	-	-	-	-	-
	MBS profile	X	-	-	-	-	-	-	-	-	-
	Transcriptional strand bias profile	X	X	X	-	-	-	-	-	-	-
	Replicative strand bias profile	X	X	X	-	-	-	-	-	-	-
	Pool samples	X	-	-	-	-	-	-	-	-	-
Signature extraction	Signature extraction (NMF)	X	X	X	-	-	-	-	X	X	-
	Signature extraction (Bayes NMF)	X	-	-	X	-	-	X	-	-	-
	Signature extraction (Lasso NMF)	-	-	-	-	-	X	-	-	-	-
	Update signature names	X	-	-	-	-	-	-	-	-	-
Signature refitting	Signature refitting	X	X	X	X	X	-	-	-	-	X
	Strict signature refitting	X	-	X	X	X	-	-	-	-	X
	Strict signature refitting (best subset)	X	-	-	-	-	-	-	-	-	X
	Bootstrapped signature refitting	X	-	-	-	-	-	-	-	-	-
	Correlation bootstrapped refitting	X	-	-	-	-	-	-	-	-	-
Signature damage analysis	Signature potential damage analysis	X	-	-	-	-	-	-	-	-	-
Signature other	Plot supported profiles / signatures	X	X	X	X	X	X	X	X	X	X
	Plot and compare supported profiles	X	X	-	-	-	-	-	-	-	-
	Signature contribution heatmap	X	X	-	-	-	-	X	X	-	-
	Signature contribution barplot	X	X	-	-	-	-	X	X	-	-
	Signature/profile similarity heatmap	X	X	-	-	-	-	-	-	X	-
	Similarity with reconstructed profile barplot	X	-	-	-	-	-	-	-	-	-
Genomic distribution	Rainfall plot	X	X	-	-	-	-	-	X	X	-
	Enrichment/depletion in genomic region	X	X	-	-	-	-	-	-	-	-
	Region specific profiles	X	-	-	-	-	-	-	-	-	-
	Region specific signatures	X	-	-	-	-	-	-	-	-	-
	Unsupervised regional similarity	X	-	-	-	-	-	-	-	-	-
Lesion segregation	Lesion segregation	X	-	-	-	-	-	-	-	-	-

Mutation matrices can be generated faster

To make MutationalPatterns (v3.4.0) scalable to large cancer datasets and suitable for interactive analysis we improved the runtime of the “mut_matrix” and “mut_matrix_stranded” functions by vectorizing them. The new functions for retrieving the mutation contexts and generating the mutation matrices have also been written in a vectorized way. As a result, these functions have O(n) or better scaling as tested on a large WGS database from the HMF (Additional file 1: Figure S8)[35].

To test their improved performance, we benchmarked the “mut_matrix” and “mut_matrix_stranded” functions on the example data provided in the previous version of MutationalPatterns (Additional file 1: Figure S9). These functions are now respectively 3.4 and 2.6 times as fast on average. In other words, a mutation matrix for 1 million SBSs can now be made in only 135 seconds on a laptop, which makes these functions suitable for large cancer datasets.

Strict signature refitting improves performance

To determine how well the strict refitting method of MutationalPatterns (v3.4.0) performs as compared to the regular method which was introduced in the original version of the package (v1.4.3), we used simulated mutation matrices. These matrices were generated by sampling trinucleotide changes of 4 different randomly selected signatures. This process was repeated 300 times per matrix, to generate 300 “samples”. Each of the samples in a matrix contained the same number of mutations per signature but was composed of different signatures. The signatures were selected from the first 30 signatures of the COSMIC signature matrix. We limited our analysis to the first 30, because these are the signatures that are most often observed in cancers and therefore more accurately resemble real-life scenarios. In addition, this approach better resembles how the package is used, because users will often fit against a limited number of signatures associated with a specific tissue. By limiting ourselves to the first 30 COSMIC signatures we also reduced overfitting. Any overfitting we observed was thus not caused by us using an unusually large signature matrix. In total we generated 4 matrices, each containing 300 samples. The number of mutations per sample was respectively 200, 400, 2000 and 4000 for the 4 different matrices.

The fraction of correctly attributed mutations to the specific signatures was increased with the strict refitting approach of MutationalPatterns (v3.4.0) as compared to "regular" or "regular_10+" refitting (Additional file 1: Figure S10a). All the tested refitting methods work better when there are more mutations per signature. Instead of using the number of correctly attributed mutations as a readout for performance, we determined whether the presence and absence of specific signatures was correctly classified. This readout might be more informative for mutational signature analysis because the presence of a signature can be a clinically relevant finding. The strict refitting method achieved a much higher precision than the original methods, while retaining a high correct recall rate (sensitivity) (Additional file 1: Figure S10b). The strict method obtained an area under the curve (AUC) of 0.925, even when only 50 mutations were present per signature, indicating that refitting can be performed on relatively small amounts of mutations.

SBS10a and SBS18 have a high damage potential

We applied the "signature_potential_damage_analysis" function on the COSMIC signatures. This analysis showed that SBS10a and SBS18 are respectively 3.6 and 2.0 times as likely to cause a "stop gain" mutation compared to a completely flat signature, containing the same weight for each mutation context, on a set of genes associated with cancer (Additional file 3: Table S2, Table S3). SBS18 is related to oxidative stress, suggesting that this type of stress has a high potency of generating premature stop codons in genes that are recurrently associated with tumorigenesis[13]. In contrast, the clock-like signature SBS1, which also occurs in healthy cells, was 0.81 and 0.40 times as likely to cause "stop gain" and "splice site" mutations, respectively, as compared to a completely flat hypothetical signature[2,36] (Additional file 3: Table S2). The damaging potential of this ageing-related mutational process is thus relatively low. Overall, C>A heavy signatures, like the recently identified ganciclovir signature, have more damage potential, because they are most likely to introduce a premature stop codon in an open reading frame[7]. Being able to quickly assess the damage potential of existing and novel signatures can be very useful to prioritize samples and mutagenic exposures for further investigation.

Applying MutationalPatterns on mutation data of DNA repair-deficiencies

To illustrate the functionality of MutationalPatterns (v3.4.0) on real-life data and to obtain novel biological insights, we applied it to mutation data obtained from cell lines in which we deleted specific DNA repair pathways using CRISPR-Cas9 genome editing technology (Additional file 1: Figure S11, Figure S12, Additional file 2). In AHH-1 cells, a lymphoblastoid cell line, we generated bi-allelic knockout lines of *MSH2*, *UNG* and *XPC* by transfecting the cells with a plasmid containing Cas9 and a single gRNA against the gene of interest. By co-transfection with a *HPRT*-targeting plasmid, we were able to select the transfected cells using 6-thioguanine, to which only *HPRT*-sufficient cells are sensitive. Using this protocol, no targeting vectors for each gene of interest were required. We analyzed somatic mutations in *HPRT*-only knockout lines as well as the combination of *HPRT* with *MSH2*, *UNG* and *XPC* (Additional file 2). To catalogue mutations that were acquired specifically in the absence of the targeted DNA repair gene, we used a previously developed method[37]. In brief, whole genome sequencing was performed on generated clones and subclones. By subtracting variants present in the clones from those in the subclones, the somatic mutations, that accumulated in between the clonal steps, were determined.

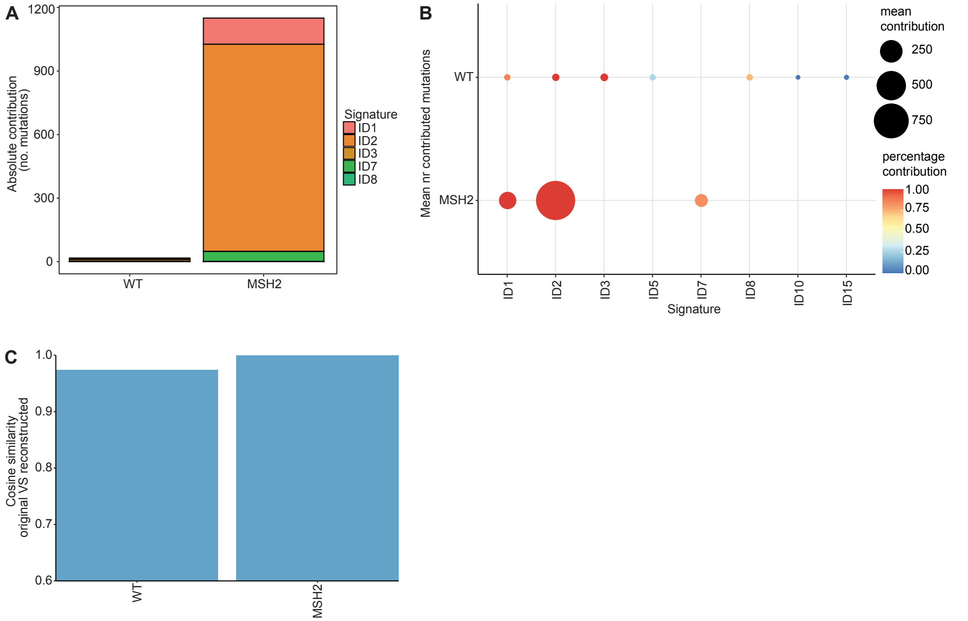


Fig. 5 Indel signatures can explain the *MSH2* profile.
a Relative contribution of each mutational signature for the wild-type (WT) and *MSH2* samples using strict signature refitting. **b** Dot plot showing the contribution of each mutational signature for the WT and *MSH2* samples using bootstrapped signature refitting. The color of a dot indicates the fraction of bootstrap iterations in which a signature contributed to a sample. The size indicates the mean number of contributing mutations across bootstrap iterations in which the contribution was not zero. **c** Bar graph depicting the cosine similarity between the original and reconstructed profiles of the WT and *MSH2* samples based on signature refitting.

The SBS profiles are shown in Additional file 1: Figure S13. Interestingly, the profile observed in the *MSH2* knockout cell line displayed a large C[C>A]T peak. When extending the sequence context surrounding the mutated base, the *MSH2* deficiency profile showed a large TT[T>C]TT peak, suggesting that this extended context surrounding mutated thymine residues is important for the underlying mutagenic process (Fig. 1d).

Next, we examined regional mutation patterns. The spectra of the *MSH2*- and *UNG*-deficient cells varied between the exonic regions and the rest of the genome, which for each sample was calculated by performing a chi-squared test using Monte Carlo simulation on a mutation spectrum matrix in which we compare the spectra of the two different regions (Fig. 2a)(fdr = 0.0012 and 0.0012, respectively). Their exons contained more C>T and less T>C mutations. The other samples did not show a significant difference in regional mutation spectra. However, when we downsampled all the samples to 227 mutations, which is the number of mutations in the *HPRT* only knockout, no significant regional mutation patterns were observed in *MSH2* and *UNG* knockout cells. This suggests that with this number of mutations insufficient statistical power was obtained for these analyses. Next to examining mutation profiles in exonic regions, we also analyzed regions with different replication timing dynamics, using the median replication timing data from 5 B-lymphocyte cell lines from ENCODE (Fig. 2b, Additional file 3: Table S4)[38]. The spectra of *MSH2* and *UNG* knockouts were different between early-, intermediate- and late-replicating DNA, which we calculated as described above (fdr = 0.0012 and fdr = 0.0012, respectively). Early replicating DNA has more C>T and less C>A than late replicating DNA. These differences were still present when downsampling was applied (fdr = 0.0025, fdr = 0.010; chi-squared test). Based on these region-specific analyses, we can conclude that the mutational processes active in the *MSH2* and *UNG* knockouts show varying activities in different regions of the genome, a result that cannot easily be obtained with other tools.

We also tested if any of the DNA repair knockout cells displayed lesion segregation, which would indicate that most of the mutations occurred during a single cell-cycle; however, this was not the case (Additional file 1: Figure S6).

Finally, we looked at the mutational signatures in the knockout samples. We performed strict signature refitting with a max_delta of 0.015 using version 3.1 of the COSMIC signatures. Signatures that were possible sequencing artefacts were excluded. Based on signature refitting, the *MSH2* knockout contained contributions of SBS5, SBS20, SBS26 and SBS44 (Fig. 3b, c). Because of the bootstrapping we can be more confident in these results. SBS5 is a clock-like signature, with unknown etiology. SBS20, SBS26 and SBS44 are all associated with defective DNA mismatch repair in cancer mutation data[13]. The *UNG* knockout contained contributions from SBS30, which has previously been attributed to deficiency of the base excision repair gene *NTHL1*[13]. The glycosylase encoded by *NTHL1* is involved in the removal of oxidized pyrimidines from the DNA and therefore SBS30 likely reflects an alternative consequence of oxidative stress-induced mutagenesis as compared to SBS18. However, *UNG* is a glycosylase that is believed to remove uracil residues from the DNA[39,40]. Therefore, our data suggests that SBS30 can be caused, besides oxidized pyrimidines, by unremoved uracil residues. Alternatively, *UNG* may also, to a certain extent, be involved in the removal of oxidized pyrimidines from the DNA. Even though the contribution of SBS30 was relatively modest in the *UNG* knockout, it was consistently picked up by the bootstrapping algorithm. This observation indicated that the number of mutations attributed to a signature is not necessarily related to the confidence of its presence, which further demonstrates the importance of our bootstrapping approach. Unexpectedly, the contribution of SBS30 in *UNG* knockout cells was negatively correlated with SBS2, even though their cosine similarity is only 0.46 (Fig. 3d). This indicates that the refitting algorithm has difficulty choosing between SBS2 and SBS30. Such difficulties in signature selection could lead to different and possibly incorrect signatures being attributed to similar sample types. Understanding the correlation of estimated signature contributions between different signatures, which can be achieved with bootstrapping, is important to prevent incorrect interpretation of the data. The *XPC* knockout contained contributions from SBS8. The etiology of this signature is not yet known. However, this finding further confirms the association of SBS8 with nucleotide excision repair deficiency[41,42]. Overall, the COSMIC signatures could explain the mutation profiles of most samples quite well, even when strict refitting was used (Fig. 3e).

Next, we studied the indel signatures in these knockout lines. Deletion of *MSH2* resulted in an increased number of indels as compared to wild-type cells (Fig. 1b). Most of these indels were single thymine deletions in thymine mononucleotide repeat regions. Signature analysis indicated that ID1, ID2 and ID7 contributed to the indel pattern in the *MSH2*-deficient cells (Fig. 5a, b). Of these, ID1 and ID2 are associated with polymerase slippage during DNA replication and found in large numbers in cancers with mismatch repair deficiency. ID7 is also associated with defective DNA mismatch repair, but not attributed to polymerase slippage[13]. Together these signatures could explain the mutational indel profile of *MSH2* knockout cells very well (Fig. 5c), showing that MutationalPatterns can perform indel signature refitting. None of the knockout cells displayed a strongly increased number of DBSs as compared to the wild-type cells (Fig. 1c).

Discussion

The novel version of MutationalPatterns (v3.4.0) has been designed to be easy-to-use in such a way that both experienced bioinformaticians and wet-lab scientists with a limited computational background can use it. The code is written in the tidyverse style, which makes it more similar to natural English and therefore easier to understand for non-programmers. MutationalPatterns (v3.4.0) gives clear error messages with tips on how to solve them, in contrast to the default error messages in R, which can sometimes be cryptic. The updated vignette, accompanying the package, not only explains how the functions in the package can be used, but also informs users on the pros and cons of the different analysis strategies.

Similar to the previous version of the package, plots are all generated using ggplot2[43]. This allows users to visualize their data in highly customizable plots that can be easily modified. Because this feature was not readily apparent for many users of the original MutationalPatterns package (v1.4.3), we have now explicitly showed how to modify the elements of a plot, such as the axis and theme, in the vignette.

We have adopted unit testing for this version of the package, resulting in more than 90% code coverage. This will improve the stability of the package and makes it easier to maintain.

The results obtained with MutationalPatterns are influenced by the quality of the variant calls that are used as its input. Since sequencing artefacts are generally not random, they can result in the detection of non-existent mutation patterns[13,44]. Therefore, users should ensure that their variant calls are stringently filtered for high-confidence variants. Additionally, since artefacts can vary based on the used sequencing techniques and bioinformatics analyses, care should be taken when comparing variant calls from different sources[45].

The novel version of MutationalPatterns (v3.4.0) is already available on Bioconductor as an update of the previous version. MutationalPatterns (v3.4.0) does not break existing scripts and pipelines, because backwards incompatible changes have been kept to a minimum.

Conclusions

MutationalPatterns (v3.4.0) is an easy-to-use R/Bioconductor package that allows in-depth analysis of a broad range of patterns in somatic mutation catalogues, supporting single and double base substitutions as well as small insertions and deletions. Here, we have described the new and improved features of the package and shown how the package performs on existing cancer data sets and on mutation data obtained from cell lines in which specific DNA repair genes are deleted. These analyses demonstrate how the package can be used to generate novel biological insights.

Mutational pattern analyses have proven to be a powerful approach to dissect mutational processes that have operated in cancer and to support treatment decision making in personalized medicine. Therefore, mutational patterns hold a great promise for improved future cancer diagnosis. The MutationalPatterns package can be used to fulfill this promise and we are confident that it will be embraced by the community.

Availability and requirements

The availability and requirements are listed as follows:
Project name: MutationalPatterns
Project home page: https://github.com/ToolsVanBox/MutationalPatterns
Archived version: https://bioconductor.org/packages/3.14/bioc/html/MutationalPatterns.html
Operating system(s): Linux, Windows or MacOS
Programming language: R (version > = 4.1.0)
License: MIT

List of abbreviations

HR: homologous recombination
Indels: Insertions and deletions
DBS: double base substitutions
VCF: variant call format
MBS: Multi base substitutions
COSMIC: Catalogue of Somatic Mutations in Cancer
NMF: non-negative matrix factorization
Bayes: Bayesian
AUC: Area under the curve
PCA: Principal component analysis
CI: Confidence interval
WT: wild-type
Mb: mega bases

Declarations

Ethics approval and consent to participate

Not applicable

Consent to publish

Not applicable

Availability of data and materials

The datasets supporting this article are available on EGA under accession number (Study ID EGAS00001004789).
Additionally, the VCF files and scripts that can be used to reproduce all figures in this paper can be found at
https://github.com/ToolsVanBox/MutationalPatterns_manuscript2_data_scripts/

Competing interests

The authors declare that they have no competing interests.

Funding

This work was financially supported by a NWO VIDI grant project 016.Vidi.171.023 to R.v.B.

Authors' contributions

F.M., R.v.B. and A.M.B wrote the manuscript. F.M. and J.d.K. developed and implemented the package. F.M. and R.O. maintain the package. A.M.B. and M.V. generated the data. F.M. and M.J.v.R. analyzed the data. A.v.H., B.v.d.R. and E.C. tested the package and provided feedback. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Francis Blokzijl and Roel Janssen for developing and maintaining the first version of this package. We also want to thank Roel Janssen for his support during the handover of the package. Finally, we would like to thank anybody who tested the package for their feedback.

References

1. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014;15:585–98.
2. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538:260–4.
3. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93.
4. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. 2013;3:246–59.
5. Pleguezuelos-Manzano C, Puschhof J, Rosendahl Huber A, van Hoeck A, Wood HM, Nomburg J, et al. Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature. 2020;580:269–73.

6. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177:821-836.e16.
7. de Kanter JK, Peci F, Bertrums E, Rosendahl Huber A, van Leeuwen A, van Roosmalen MJ, et al. Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. Cell Stem Cell. 2021
8. Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat Med. 2017;23:517–25.
9. Nguyen L, W. M. Martens J, Van Hoeck A, Cuppen E. Pan-cancer landscape of homologous recombination deficiency. Nat Commun. 2020;11:5584.
10. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, Lopez E, et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature. 2005;434:913–7.
11. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, et al. Inhibition of Poly(ADP-Ribose) Polymerase in Tumors from BRCA Mutation Carriers. N Engl J Med. 2009;361:557–68.
12. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 2018;
13. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
14. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016;17:31.
15. Ramazzotti D, Lal A, Liu K, Tibshirani R, Sidow A. De Novo Mutational Signature Discovery in Tumor Genomes using SparseSignatures. bioRxiv. 2019;384834.
16. Rosales RA, Drummond RD, Valieris R, Dias-Neto E, Da Silva IT. signeR: An empirical Bayesian approach to mutational signature discovery. Bioinformatics. 2017;33:8–16.
17. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. Bioinformatics. 2015;31:3673–5.
18. Mayakonda A, Lin D-C, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018/10/19. 2018 Nov;28:1747–56.
19. Krüger S, Piro RM. decompTumor2Sig: identification of mutational signatures active in individual tumors. BMC Bioinformatics. 2019;20(4):152.
20. Maura F, Degasperi A, Nadeu F, Leongamornlert D, Davies H, Moore L, et al. A practical guide for mutational signature analysis in hematological malignancies. Nat Commun. 2019;10:2969.
21. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518:360–4.
22. Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science (80-). 2019;364:eaaw2872.
23. Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. Local Determinants of the Mutational Landscape of the Human Genome. Cell. 2019;177:101–14.
24. Aitken SJ, Anderson CJ, Connor F, Pich O, Sundaram V, Feig C, et al. Pervasive lesion segregation shapes cancer genome evolution. Nature. 2020;583:265–70.
25. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv. 2019;861054.
26. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995 Dec 17;57(1):289–300.
27. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;48:D682–8.
28. Lawson CL, Hanson RJ. Solving least squares problems. SIAM; 1995.
29. Woo J, Winterhoff BJ, Starr TK, Aliferis C, Wang J. De novo prediction of cell-type complexity in single-cell RNA-seq and tumor microenvironments. Life Sci Alliance. 2019;2:e201900443.
30. Degasperi A, Amarante TD, Czarnecki J, Shooter S, Zou X, Glodzik D, et al. A practical framework

and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. Nat cancer. 2020;1:249–63.

31. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. Genome Biol. 2015 Mar;16:56.

32. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018 Feb;

33. Chi X, Li Y, Qiu X. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. Immunology. 2020/02/27. 2020 Jul;160(3):233–47.

34. Di Noia JM, Neuberger MS. Molecular Mechanisms of Antibody Somatic Hypermutation. Annu Rev Biochem. 2007 Jun 7;76(1):1–22.

35. Priestley P, Baber J, Lolkema MP, Steeghs N, de Bruijn E, Shale C, et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature. 2019;575:210–6.

36. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet. 2015;47:1402–7.

37. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science (80-). 2017;238:eaao3130.

38. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710.

39. Prasad A, Wallace SS, Pederson DS. Initiation of Base Excision Repair of Oxidative Lesions in Nucleosomes by the Human, Bifunctional DNA Glycosylase NTH1. Mol Cell Biol. 2007;27:8442 LP – 8453.

40. Li J, Braganza A, Sobol RW. Base Excision Repair Facilitates a Functional Relationship Between Guanine Oxidation and Histone Demethylation. Antioxid Redox Signal. 2013;18:2429–43.

41. Jager M, Blokzijl F, Kuijk E, Bertl J, Vougioukalaki M, Janssen R, et al. Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. Genome Res. 2019;29:1067–77.

42. Yurchenko AA, Padioleau I, Matkarimov BT, Soulier J, Sarasin A, Nikolaev S. XPC deficiency increases risk of hematologic malignancies through mutator phenotype and characteristic mutational signature. Nat Commun. 2020;11(1):5834.

43. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.

44. Alexandrov LB, Nik-zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer Tumor Cells Carry Somatic Mutations. CellReports. 2012;3(1):246–59.

45. Chen Z, Yuan Y, Chen X, Chen J, Lin S, Li X, et al. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. Sci Rep 2020;10(1):3501.

46. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2021 Dec 1;gkab1112.

Supplementary material

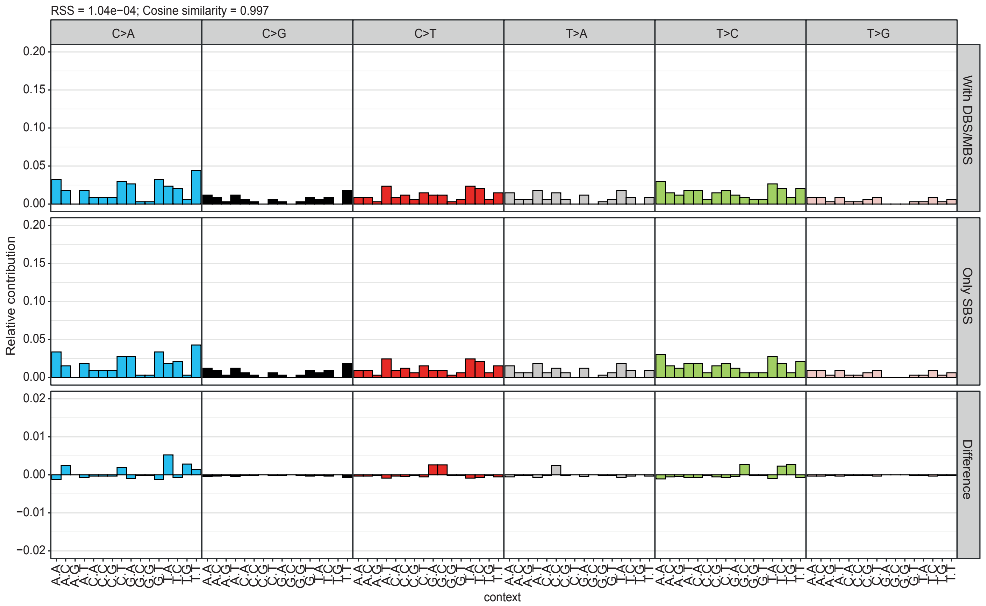


Fig. S1 Removing the DBSs and MBSs results in an improved SBS profile.
Relative contribution of each of the 96 trinucleotide changes to the mutational profiles of the wild-type sample. The upper panel shows the profile when DBSs and MBSs are incorrectly classified as SBSs. The middle panel shows the profile with only the SBSs. The lower panel shows the difference between these profiles.

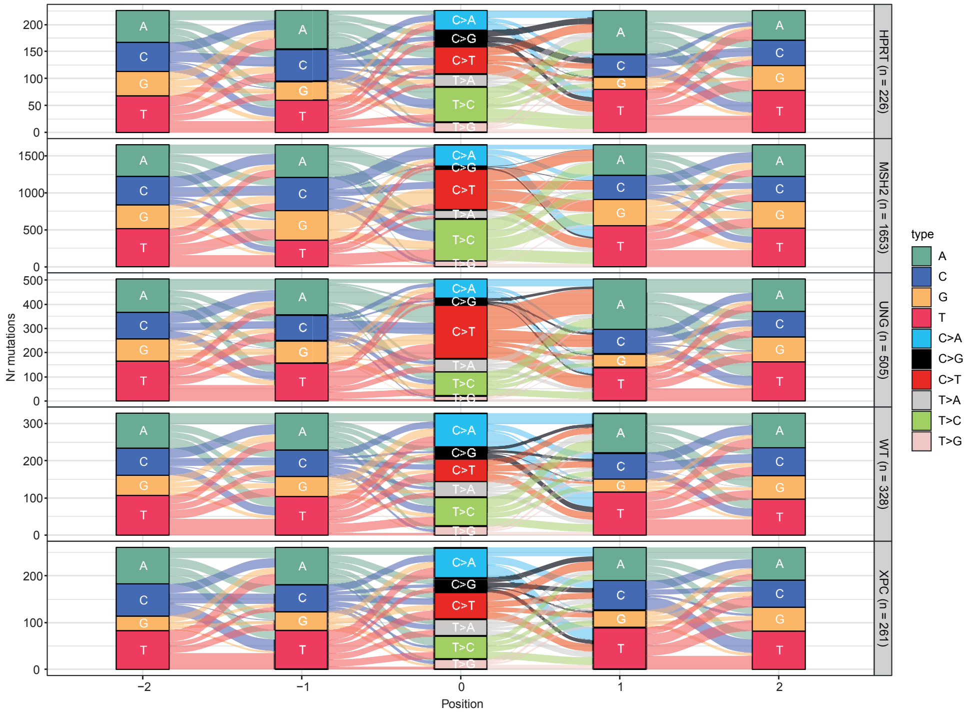


Fig. S2 Mutation contexts can be visualized with a river plot.

A river plot depicting the indicated mutation types and the surrounding context for each sample. The bars show the number of mutations or bases for each type. The flows show the connections between the mutations and their context.

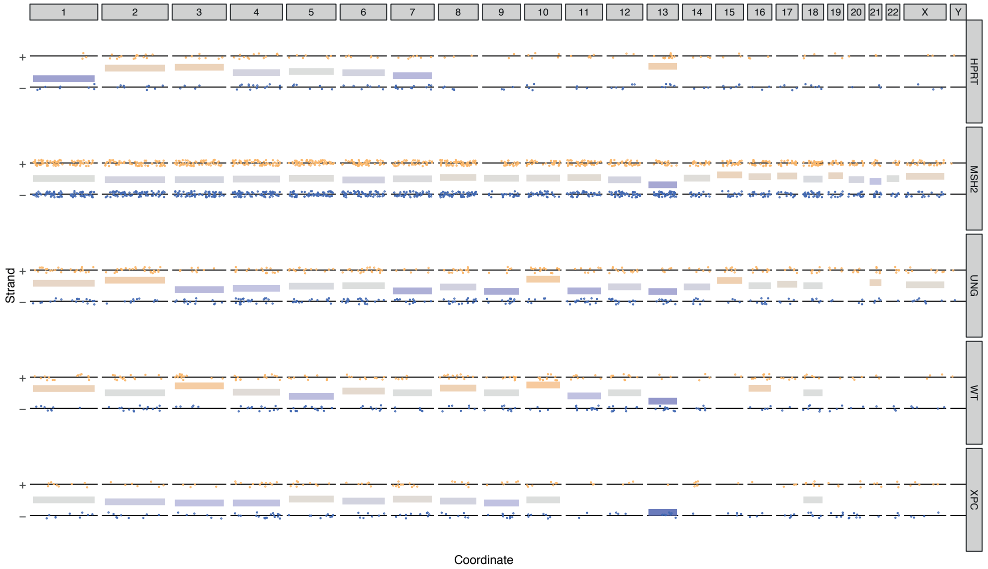


Fig. S3 Lesion segregation can be visualized.
A jitter plot depicting the presence of lesion segregation for each sample per chromosome. Each dot depicts a single base substitution. Any C>N or T>N is shown as a "+" strand mutation, while G>N and A>N mutations are shown on the "-" strand. The x-axis shows the position of the mutations. The horizontal lines are calculated as the mean of the "+" and "-" strand, where "+" equals 1 and "-" equals 0. They indicate per chromosome on which strand most of the mutations are located. In this example no lesion segregation was present.

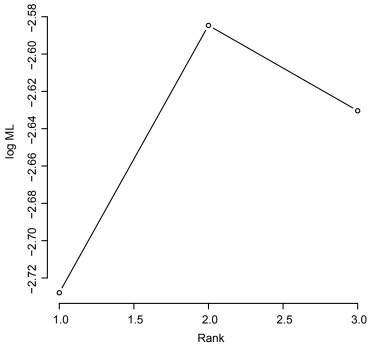


Fig. S4 Variational Bayes NMF can be used to predict the optimal number of signatures to extract.
The log maximum likelihood is shown for different ranks. The rank is the number of signatures to extract. The highest likelihood shows the optimal rank. In this case this is 2.

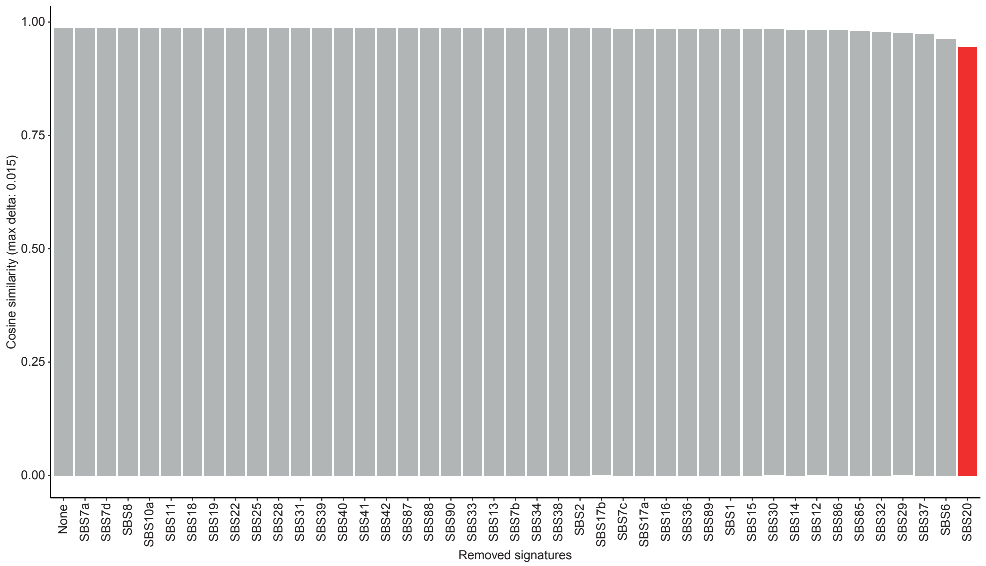


Fig. S5 Strict refitting iteratively removes signatures.
The cosine similarity between the original and reconstructed profile of the MSH2 knockout during different iterations of the strict refitting process. The signatures with the lowest contributions are iteratively removed and the cosine similarity is calculated. This is depicted from left to right. Removing SBS20 decreased the cosine similarity more than the cutoff, so it was retained, and the algorithm stopped.

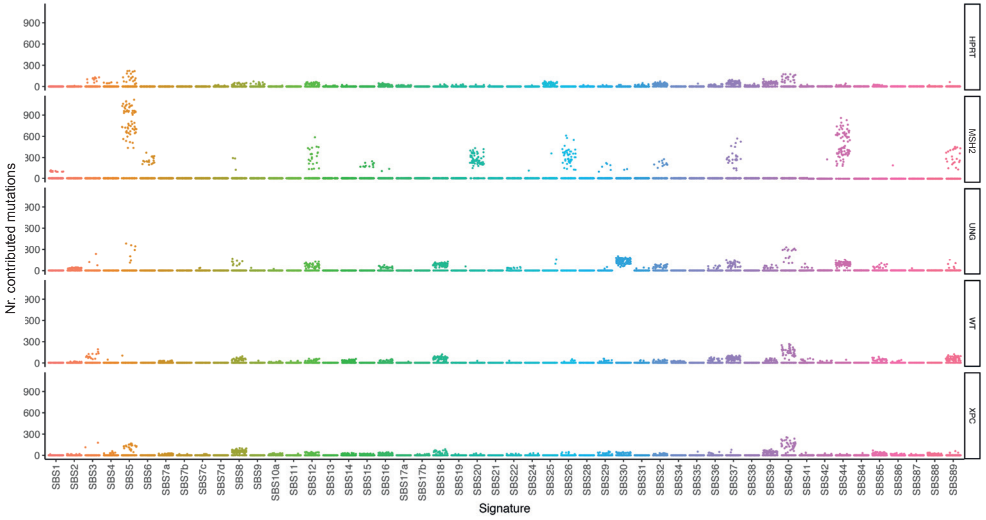


Fig. S6 Bootstrapped signature refitting can be visualized with a jitter plot.
A jitter plot depicting the bootstrapped signature refitting for each sample. Each dot shows the number of mutations contributed by a signature according to one bootstrap iteration.

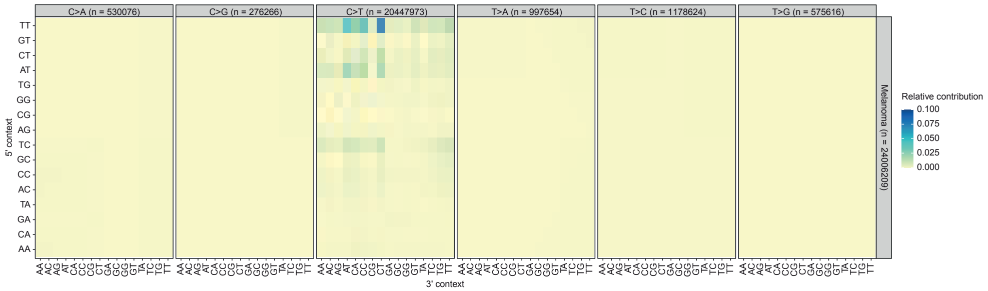


Fig. S7 TT[C>T]CT is the most common substitution type in metastatic melanomas.
Heatmap depicting the relative contribution of the indicated mutation types and the surrounding bases to the point mutation spectrum for metastatic melanomas. The total number of somatic point mutations is indicated. In contrast to Fig. 4a, all substitutions are shown.

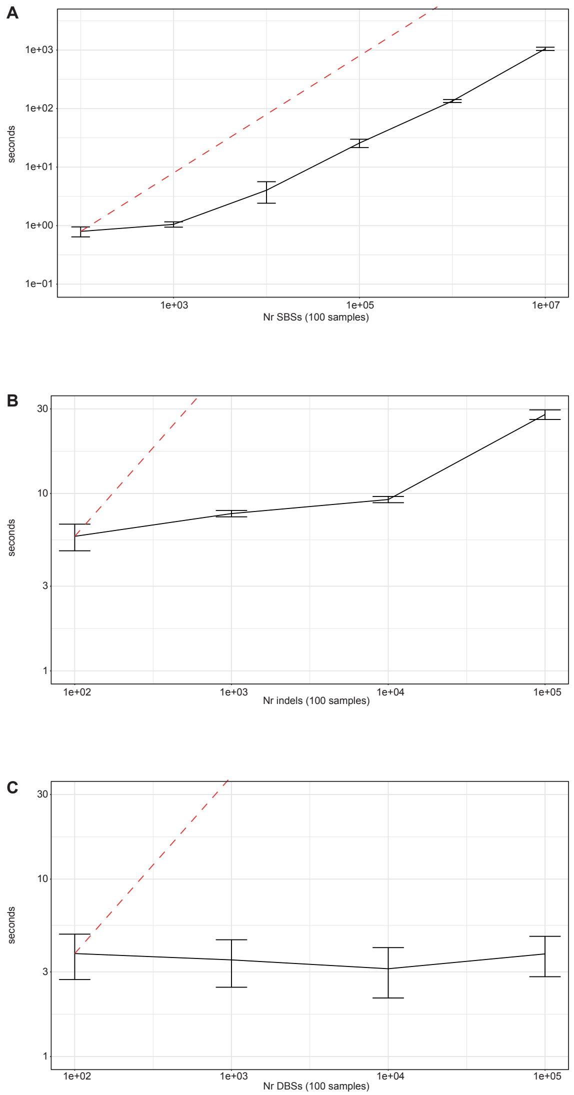


Fig. S8 The matrix generating functions have O(n) or better scaling.
The y-axis shows the time it takes to generate a mutation matrix for the number of mutations on the x-axis for **a** SBSs, **b** indels and **c** DBSs. The mutations are always split over 100 samples. The dashed red line indicates O(n) scaling. The SBS and indel functions approach O(n) scaling on large mutations sets. The runtime of the DBS function is independent of the number of mutations.

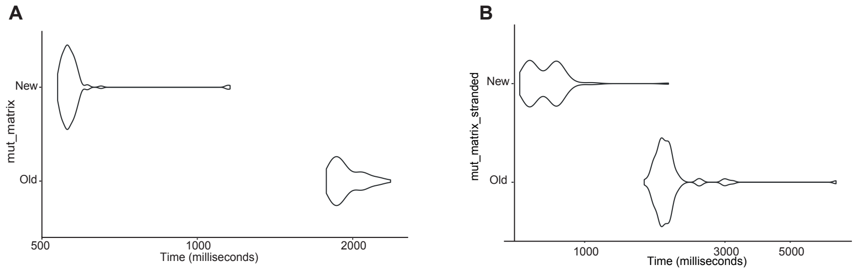


Fig. S9 Benchmark of the “mut_matrix” function.
Violin plot depicting the run-times of the new and the old versions of the **a** “mut_matrix” and **b** “mut_matrix_stranded” functions. The benchmark was run on a 2019 MacBook Pro (2.4 GHz Quad-Core, 16GB RAM).

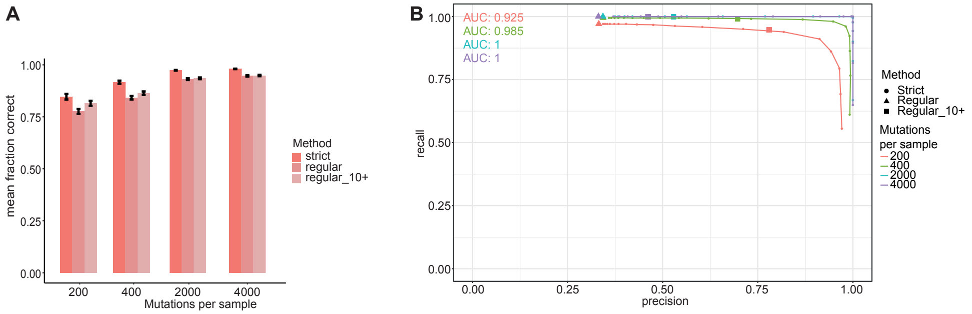


Fig. S10 Recall-precision plot of different refitting methods in MutationalPatterns.
a Bar graph depicting the mean fraction of correctly attributed mutations for “regular”, “regular_10+” and “strict” refitting. This is shown for 4 experiments. Per experiment a mutation matrix with 300 simulated samples, each containing 4 signatures, was generated. The number of mutations per sample was respectively 200, 400, 2000 and 4000 for the 4 different experiments. The error bars show the 95% confidence interval. The fraction of correctly attributed mutations is calculated as 1 minus the absolute difference between the real and estimated contribution divided by the sum of the real and estimated contribution. **b** Recall-precision plot showing the recall (sensitivity) and precision of the “strict” method when different “max_delta” cutoffs are used for signature refitting. The recall and precision of the “regular” and “regular_10+” methods are also shown with respectively triangles and squares. Since these methods don’t have a “max_delta” cutoff only a single point can be shown for them. This is shown for 4 experiments. Per experiment a mutation matrix with 300 simulated samples, each containing 4 signatures, was generated. The number of mutations per sample was respectively 200, 400, 2000 and 4000 for the 4 different experiments. The area under the curve (AUC) is shown per experiment.

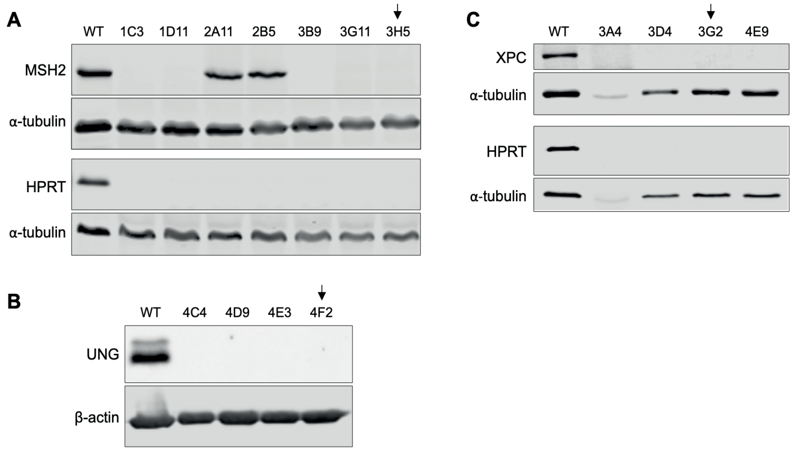


Fig. S11 Western blot analysis of AHH-1 CRISPR-Cas9 edited clonal lines.
a Western blot of total protein lysate of bulk AHH-1 (WT) or single cell clones generated after transfection with CRISPR-Cas9 plasmids targeting *MSH2* and *HPRT*. All clones are *HPRT* knockout, and 5/7 clones are also knockout for *MSH2*. **α**-Tubulin staining was done on the same membrane as the indicated protein above. **b** Western blot of total protein lysate of bulk AHH-1 (WT) or single cell clones generated after transfection with CRISPR-Cas9 plasmids targeting *UNG* and *HPRT*. All clones are knockout for UNG and HPRT (HPRT blot not shown). **ß**-actin staining was done on the same membrane as UNG staining. **c** Western blot of total protein lysate of bulk AHH-1 (WT) or single cell clones generated after transfection with CRISPR-Cas9 plasmids targeting *XPC* and *HPRT*. All clones are *HPRT* and *MSH2* knockout. **α**-Tubulin staining was done on the same membrane as the indicated protein above. Arrows indicate clones selected for a second clonal step and whole genome sequencing. Full-length blots/gels are presented in Additional file 1: Figure S12.

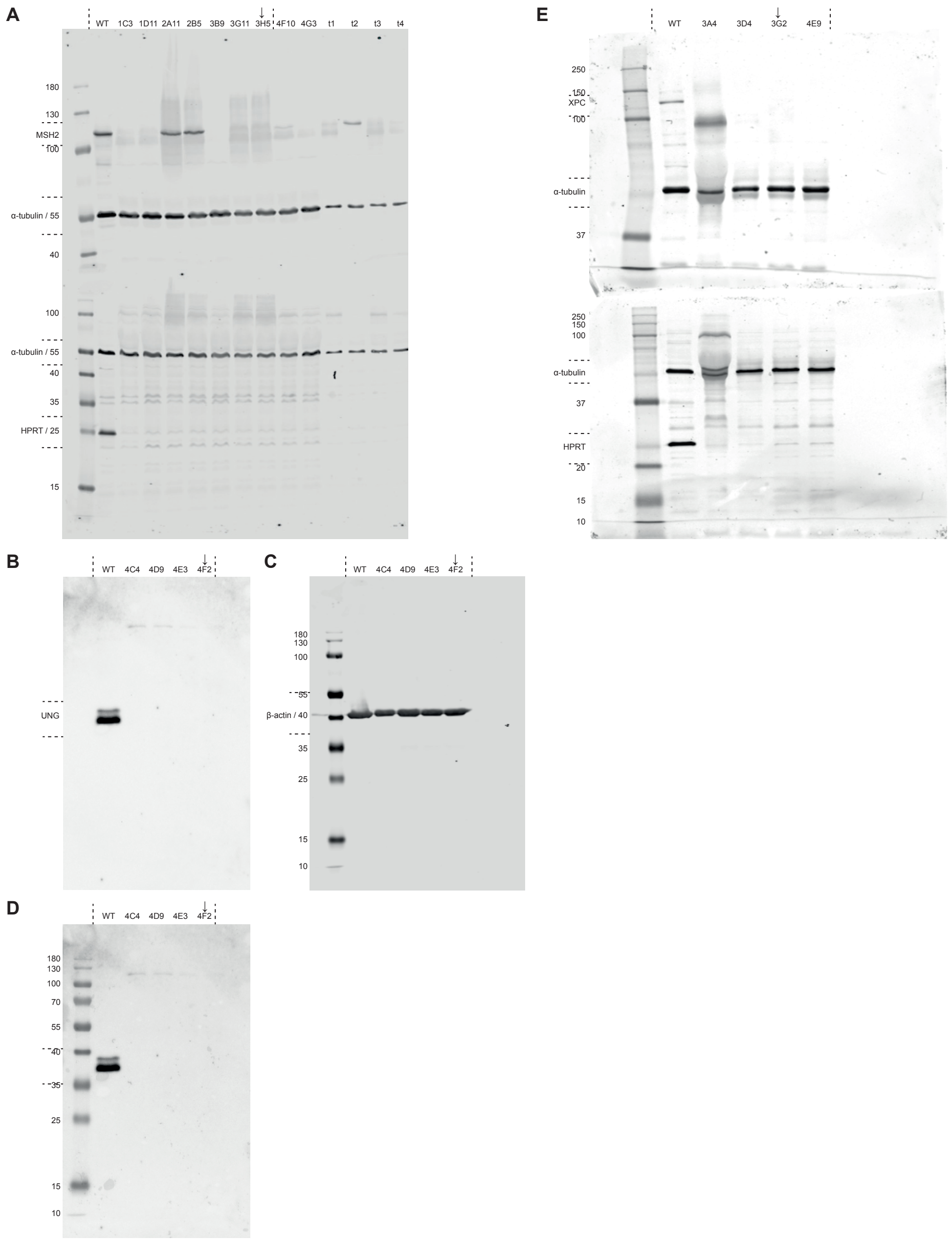


Fig. S12 Uncropped original version of the western blots in Fig. S11.
The vertical and horizontal dashed lines indicate the crop marks. Standard protein size markers have been labeled with the expected molecular weight in kDa. **a** Corresponds to Fig. S11a. **b**, **c** and **d** correspond to Fig. S11b. The UNG image was developed using ECL (b), whereas the b-actin and protein ladder were imaged using fluorescent antibodies (c). A composite image of (b) is shown in (d), which includes the brightfield image of the same membrane that contains the protein ladder. **e** Corresponds to Fig. S11c. PageRuler Prestained Protein Ladder was used for (a), (b), (c), and (d). Precision Plus Protein WesternC Standards was used for (e).

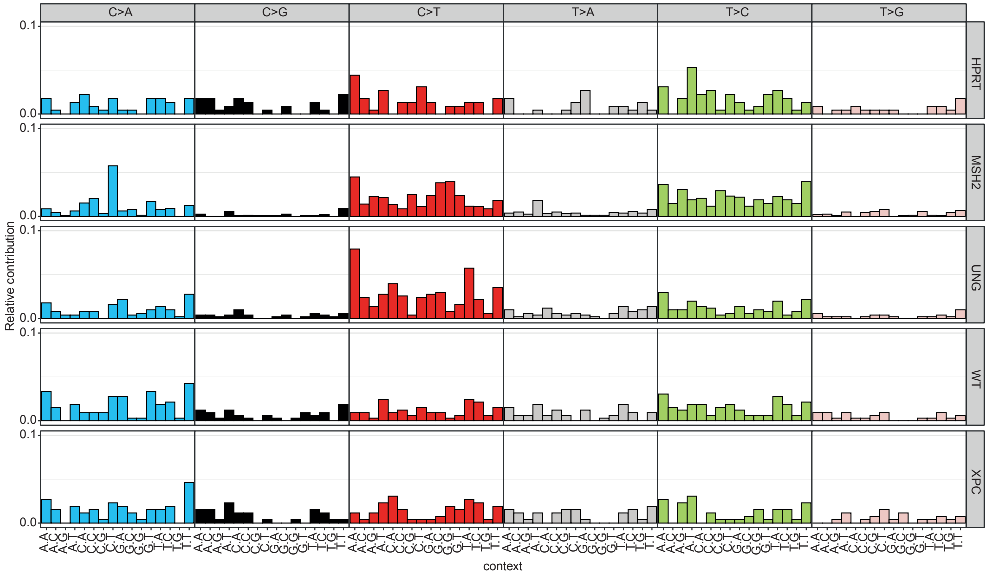


Fig. S13 SBS profiles of knockout samples.
Relative contribution of each trinucleotide change to the point mutation spectrum for each sample.

For Table S1-S4 see: https://doi.org/10.1186/s12864-022-08357-3

Additional methods

Cell lines

AHH-1 cells, a human B lymphoblastoid cell line, were obtained from ATCC. Cells were cultured in RPMI 1640 medium (Gibco) with 2 mM L-glutamine, 1.5 g/L sodium bicarbonate, 4.5 g/L glucose, 10 mM HEPES, 1.0 mM sodium pyruvate, 10% heat-inactivated horse serum and penicillin/streptomycin. Low-passage cells were used for transfection experiments.

CRISPR-Cas9 gene editing

The human codon-optimized Cas9 expression plasmid was obtained from Addgene (px330-U6-Chimeric_BB-CBh-hSpCas9). The gRNA sequences were inserted by BbsI digestion and T4 ligation as described[1]. sgRNA target sequences: hHPRT-sgRNA 5'-GGCTTATATCCAACACTTCG-3', hMSH2-sgRNA 5'-ACAAAGACTTGTTAACCAG-3', hUNG-sgRNA 5'-TCGGCACTCAGCGGCGAGGA-3', hXPC-sgRNA 5'-AAAGATTGACTGCGGATCC-3'. To generate DNA repair gene knockout lines, single cell suspensions of AHH-1 cells were co-transfected with Cas9- and sgRNA-expressing px330 plasmids, targeting HPRT and either MSH2, UNG or XPC. Plasmid DNA was mixed in an equal ratio and combined with Lipofectamine 2000. Transfection was performed according to manufacturer's instructions in AHH-1 medium with 1% horse serum. After 3 hours, complete AHH-1 medium (10% horse serum) was added to the cells and cells were cultured for 6 days. On day 6, 0.5 µg/mL 6-thioguanine (6-TG) was added to the cells after making single cell suspensions to select for HPRT knockout cells. On day 19, cells growing on 6-TG were plated in limiting dilutions to obtain single cell clones. Growing clones were analyzed by PCR and Western blot for knockout of HPRT and MSH2, UNG or XPC. Selected clones were subjected to another round of limiting dilutions to obtain subclones of the selected clones.

Western blot

Cell pellets were directly lysed in sample buffer (62.5 mM Tris-HCl, 2.5% SDS, 10% glycerol, 0.002% bromophenol blue, 100 mM DTT). Total protein lysates were loaded on SDS-PAGE gels (XPC + MSH2: 8%, UNG + HPRT: 12%) and transferred to nitrocellulose membranes (BioRad). Membranes were blocked and probed with antibodies directed against HPRT (ab10479, Abcam), MSH2 (D24B5, Cell Signaling Technology), UNG (OTI1A11, ThermoFisher Scientific), XPC (12701, Cell Signaling Technology), **ß**-actin (RM112, Sigma-Aldrich) and **α**-tubulin (B-5-1-2, Sigma-Aldrich). PageRuler Prestained Protein Ladder (26617, Thermo Scientific) and Precision Plus Protein WesternC Standards (5561, Bio-Rad) were used as protein standards, where indicated. Membranes stained by fluorescent antibodies were scanned using the Odyssey CLx (LI-COR). Images were exported using Image Studio 5.2 (LI-COR). Membranes stained by ECL antibodies were developed using Pierce™ ECL Western Blotting Substrate (32106, Thermo Scientific) and scanned using the ChemiDoc™ Touch Imaging System (Bio-Rad). ECL images were processed using the on-board Image Lab. All image processing was restricted to linear adjustments (brightness, contrast) to visualize the bands. Subsequently, colored images were converted to greyscale using ImageJ 2.0.0-rc-69/1.52i.

Whole genome sequencing and read alignment

DNA libraries for Illumina sequencing were generated by using standard protocols (Illumina) from 500 ng of genomic DNA isolated from the clonally expanded AHH-1 cells using QIAamp DNA Blood & Tissue Kit (QIAGEN) according to manufacturers' instructions. All samples were sequenced (2 × 150 bp) by using Illumina HiSeq X Ten or NovaSeq 6000 sequencers to 30X base coverage. Whole genome sequencing data was mapped against human reference genome GRCh38 by using Burrows-Wheeler Aligner v0.7.5a mapping tool[2] with settings 'bwa mem -c 100 -M'. Sequence reads were marked for duplicates by using Sambamba v0.6.8 markdup. Full pipeline description and settings also available at: https://github.com/UMCUGenetics/IAP.

Mutation calling and filtering

Raw variants were multisample-called by using the GATK HaplotypeCaller v3.8-1-0[3] and GATK-Queue v3.8-1-0 with default settings and additional option 'EMIT_ALL_CONFIDENT_SITES'. The quality of variant and reference positions was evaluated by using GATK VariantFiltration v3.8-1-0 with options -snpFilterName SNP_LowQualityDepth -snpFilterExpression "QD < 2.0" -snpFilterName SNP_MappingQuality -snpFilterExpression "MQ < 40.0" -snpFilterName SNP_StrandBias -snpFilterExpression "FS > 60.0" -snpFilterName SNP_HaplotypeScoreHigh -snpFilterExpression "HaplotypeScore > 13.0" -snpFilterName SNP_MQRankSumLow -snpFilterExpression "MQRankSum < -12.5" -snpFilterName SNP_ReadPosRankSumLow -snpFilterExpression "ReadPosRankSum < -8.0" -snpFilterName SNP_HardToValidate -snpFilterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)" -snpFilterName SNP_LowCoverage -snpFilterExpression "DP < 5" -snpFilterName SNP_VeryLowQual -snpFilterExpression "QUAL < 30" -snpFilterName SNP_LowQual -snpFilterExpression "QUAL >= 30.0 && QUAL < 50.0 " -snpFilterName SNP_SOR -snpFilterExpression "SOR > 4.0" -cluster 3 -window 10 -indelType INDEL -indelType MIXED -indelFilterName INDEL_LowQualityDepth -indelFilterExpression "QD < 2.0" -indelFilterName INDEL_StrandBias -indelFilterExpression "FS > 200.0" -indelFilterName INDEL_ReadPosRankSumLow -indelFilterExpression "ReadPosRankSum < -20.0" -indelFilterName INDEL_HardToValidate -indelFilterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)" -indelFilterName INDEL_LowCoverage -indelFilterExpression "DP < 5" -indelFilterName INDEL_VeryLowQual -indelFilterExpression "QUAL < 30.0" -indelFilterName INDEL_LowQual -indelFilterExpression "QUAL >= 30.0 && QUAL < 50.0" -indelFilterName INDEL_SOR -indelFilterExpression "SOR > 10.0". To obtain high-quality somatic mutation catalogs, we applied postprocessing filters as described[4]. Briefly, we considered variants at autosomal chromosomes without any evidence from a paired control sample (the original bulk culture used to generate the mutant lines); passed by VariantFiltration with a GATK phred-scaled quality score R 100; a base coverage of at least 10X in the clonal and subclonal cultures, and paired control sample; mapping quality (MQ) of 60; no overlap with single nucleotide polymorphisms (SNPs) in the Single Nucleotide Polymorphism Database v146; and absence of the variant in a panel of unmatched normal human genomes (BED-file available upon request). We additionally filtered heterozygous base substitutions with a GATK genotype score (GQ) lower than 99 in clonal or paired control samples. A GQ score of 10 was used for homozygous variants. For indels, we filtered variants with a GQ score lower than 99 in both clonal or subclonal culture, or paired control sample. A GQ of 20 was used for homozygous reference variants[4,5]. Finally, we only considered variants with a variant allele frequency of ≥0.3 in the sub-clones and a variant allele frequency lower than 0.3 in the original paired clones. These variants specifically accumulated between the two clonal expansion steps. The script is available at: https://github.com/ToolsVanBox/SMuRF.

Additional references

1. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. 8, 2281–2308.
2. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595.
3. Depristo, M.A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–501.
4. Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. Nature 538, 260–264.
5. Jager, M., Blokzijl, F., Sasselli, V., Boymans, S., Janssen, R., Besselink, N., Clevers, H., van Boxtel, R., and Cuppen, E. (2018). Measuring mutation accumulation in single human adult stem cells by whole-genome sequencing of organoid cultures. Nat. Protoc. 13, 59–78.

Chapter 4

Mutation accumulation in mitochondrial DNA of cancers resembles mutagenesis in normal stem cells

Freek Manders[1], Jip van Dinter[1], and Ruben van Boxtel[1]

[1]Princess Máxima Center for Pediatric Oncology and Oncode Institute, Heidelberglaan 25,
3584CS Utrecht, The Netherlands

iScience, 2022

Summary

Mitochondria are small organelles that play an essential role in the energy production of eukaryotic cells. Defects in their genomes are associated with diseases such as cancer, as well as aging. Here we analyzed the mitochondrial genomes of 532 whole genome sequencing samples from cancers and normal clonally expanded single cells. We show that the mitochondria of normal cells accumulate mutations with age and that most of the mitochondrial mutations found in cancer are the result of healthy mutation accumulation. We also show that the normal HSPCs of leukemia patients have an increased mitochondrial mutation load. Finally, we show that secondary pediatric cancers and chemotherapy treatments do not impact the mitochondrial mutation load and mitochondrial DNA copy numbers of most cells, suggesting that damage to the mitochondrial genome is not a major driver for carcinogenesis. Overall, these findings may contribute to our understanding of mitochondrial genomes and their role in cancer.

Keywords

Mitochondria, Whole genome sequencing, somatic mutations, copy numbers, mutational processes, Hematopoietic stem cells, leukemia, single cells

Introduction

Mitochondria, known as “the powerhouses of the cell”, are small organelles that play an essential role in the energy production of eukaryotic cells. They also play a role in many other cellular processes, such as apoptosis, biosynthesis and cellular differentiation[1–3]. A single cell harbors many mitochondria, ranging from a couple dozen to more than a thousand depending on the cell type[4,5]. Mitochondria contain their own mitochondrial DNA (mtDNA), of which up to 15 copies can be present per mitochondrion[6]. The mtDNA is circular and, even though it is only 16.6kb, contains 37 genes of which 13 are protein coding. The remaining 24 genes, consisting of 22 tRNAs and 2 ribosomal RNAs, are used for translation of the 13 protein coding genes. Unlike in nuclear DNA, mitochondrial genes lack introns or non-coding intergenic sequences[7]. Genetic variation in the mtDNA of a cell is often present in only a subset of its mitochondria, a phenomenon known as heteroplasmy[8].

Defects in the mitochondrial genome have been associated with the development of a variety of neurodegenerative diseases as well as aging[4,9–11]. In addition, mutations in mtDNA have also been suspected to play a role in the onset or progression of cancer[12,13]. A better understanding of mitochondrial mutations and their role in mitochondrial dysfunction is thus important to better understand cancer and other diseases. However, even though their relevance to disease is clear, mitochondrial genomes have been studied less than their nuclear counterparts and mitochondrial reads are often discarded in whole genome sequencing (WGS) studies.

Recently, the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium aggregated WGS data from 2,658 cancers, characterized mutation patterns in the mtDNA and found that the copy number of the mitochondrial genome varied greatly within and across 38 tumor types[12]. However, it remains unclear if these mutations and copy number differences precede carcinogenesis and are already present in normal tissues or if they are a consequence of malignant transformation. We and others have previously shown that normal stem cells of various tissues accumulate mutations in their nuclear genome in a linear fashion with age[14–17]. Additionally, several studies have found correlations between age and mitochondrial mutation burden in cancer, brain and colon samples, however these studies focused on bulk tissues or also included germline variants[12,18–23]. Single stem cells of normal tissues have not yet received as much attention.

Characterizing the mitochondrial genomes of single cells or clonally expanded cells is necessary to identify mutations in normal tissues. Additionally, it allows for the direct comparison of normal tissues against cancers, which are also clonal expansions of a parental malignant cell.

Here, we analyzed the mitochondrial genomes of 532 WGS samples from 88 donors[15,17,24–27]. We show that the mitochondria of normal stem cells, across different tissue types, gradually accumulate mutations with age and that most of the mtDNA mutations present in cancer occurred before transformation. Surprisingly, mitochondrial genomes are relatively insensitive to disease conditions, such as cancer, and/or treatment perturbations, such as chemotherapy. Overall, our study provides insight into mitochondrial mutation accumulation and its perturbation by cancer and cancer treatments.

Results

Cataloguing somatic mitochondrial mutations

A mitochondria specific sequencing analysis pipeline was recently developed as a part of the genome analysis toolkit (GATK)[28,29]. We applied this pipeline on samples from clonally expanded stem cells (hereafter named "clones") of normal human tissues as well as cancer samples and bulk control samples (Methods).

After removing 4 samples because of their low quality (Methods), we were left with 532 samples, originating from 88 different donors ranging in age from 0 to 87 years (Fig. 1). The average sequencing depth of the mitochondrial genome was 7250x (3788-9569 Interquartile range (IQR)), allowing for the detection of variants with a very low variant allele frequency (VAF). Almost the entire mitochondrial genome had a high read coverage allowing for the detection of somatic variants across the mitochondrial genome (Fig. S1a). The distribution of reads across the genome was also highly similar between samples with a median cosine similarity of 0.998 [range: 0.981-1.000].

In total, we identified 370 somatic mitochondrial substitutions and 33 indels with a median VAF of respectively 0.0522 (0.0240-0.1356 IQR) and 0.0295 (0.0201-0.0809 IQR) (Fig. S1b, c). Even though the median VAF was very low, the median number of reads supporting a variant was respectively 342 (150-838 IQR) and 171 (119-432 IQR). This observation indicates that these variants are unlikely to be stochastic sequencing artifacts or false positives caused by Nuclear Mitochondrial sequences (NuMTs), which are parts of the mitochondrial genome that have been inserted into the nuclear genome[30] (Fig. S1d, e). Since we identified only a limited number of indels, we focused our subsequent analyses on the base substitutions.

MT-ND5 was the most commonly mutated gene, in line with previous observations[12] however, after correcting for gene length it was no longer enriched (Fig. S1f). Most of

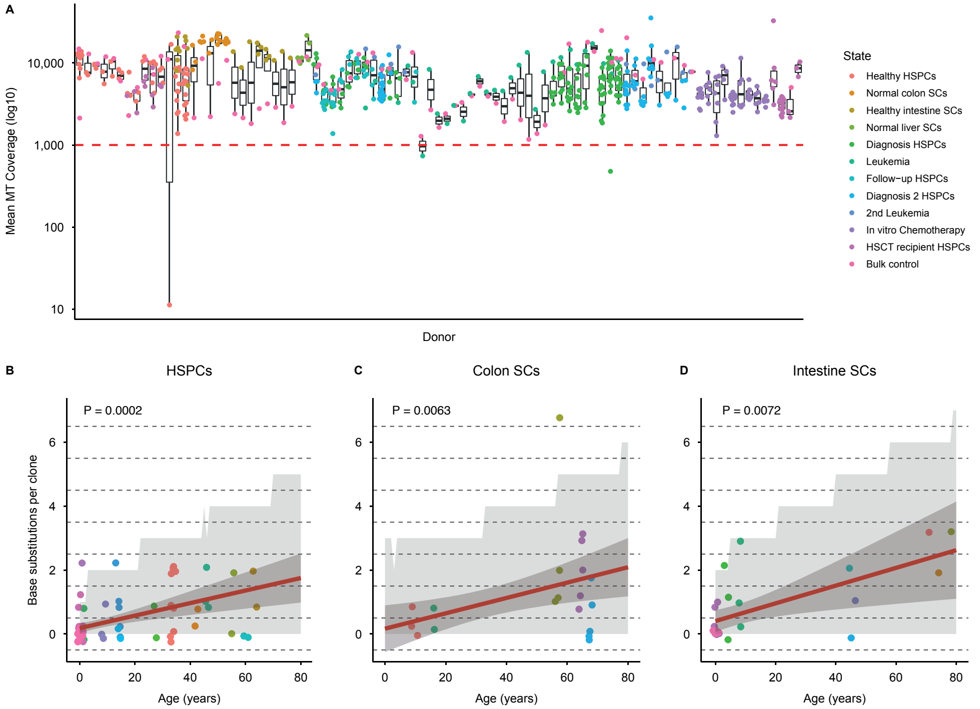


Figure 1 Accumulation of mitochondrial substitutions in normal stem cells with age.
a The mitochondrial read coverage is shown per donor, with each dot showing a single sample (536 samples; 89 donors). The color of the dots indicates the sample group. Samples below the dashed red line were removed for having a low mitochondrial read coverage. Donors were ordered on the x-axis based on the sample groups of their samples. **b**, **c**, and **d** The number of mitochondrial base substitutions per clone is plotted against the donor age for normal HSPCS (blood stem cells) (b) (33 mutations; 62 samples; 14 donors), normal colon stem cells (SCs) (c) (26 mutations; 19 samples; 5 donors), and normal intestinal stem cells (d) (20 mutations; 22 samples; 10 donors). Each clone is a clonally expanded single-cell. *p*-values show the significance of the age of the donor on the number of substitutions (generalized linear model). The red line indicates the mean fitted number of mutations at that age. The dark grey background shows the 95% confidence interval of the model, whereas the light grey background shows the 95% prediction interval. The prediction intervals show the predicted intervals that contain the mutation load of 95% of all cells in the population. A small amount of jitter was added to the dots to prevent them from completely overlapping. The color of the dots indicates the donor.

the genic substitutions were predicted to have a low to moderate effect, suggesting that they are unlikely to have a large physiological effect (Fig. S1g).

Mutation accumulation in mitochondria of normal cells

To determine the relation between mtDNA mutation burden and age, we regressed the number of base substitutions per stem cell clone against the age of the donor. In hematopoietic stem and progenitor cells (HSPCs) from healthy donors, we observed a mutation rate of 0.0196 substitutions per stem cell per year (95% confidence interval: 0.0093-0.0299; $p = 0.0002$; generalized linear model; Fig. 1b). The mutation load of HSPCs did have a weak correlation between mtDNA copy numbers and age ($p = 0.0451$; generalized linear model; Fig. S1h), similar to previous findings[12]. Samples with a higher-than-average mutation burden in their mitochondrial genome, did not have a higher-than-average burden in their nuclear genome, suggesting that for normal cells, the mutational load in the mitochondria is independent of the mutation load in the nucleus ($p = 0.726$; $X^2 = 0.313$; Chi-squared test; Fig. S1i). In normal colon and intestinal stem cells, we observed mutation rates of 0.0240 and 0.0278 substitutions per year, confirming previous results in colon (95% confidence interval: 0.0068-0.0413, 0.0075-0.0480; $p = 0.0063$, $p = 0.0072$; generalized linear model; Fig. 1c, d)[21]. The rates in colon and intestinal stem cells are not significantly different from HSPCs ($p = 0.7834$; $p = 0.1138$; generalized linear model). Additionally, the rates we found are similar to the rate of 0.0067 previously observed in human putamen, which is a part of the brain[18]. Overall, our data shows that mitochondria in stem cells gradually accumulate mutations with age in multiple tissues at comparable rates.

Mitochondrial DNA copy numbers differ between tissues

We did not observe a significant relation between mtDNA copy numbers and age in any of the studied tissues (Blood: $p = 0.3466$, Colon: $p = 0.3099$, Intestine: $p = 0.0883$; linear mixed-effects model; Fig. 2). Merging the data of these tissues to maximize statistical power did not change this result ($p = 0.3108$; linear mixed-effects model). This observation is surprising because correlations between mtDNA copy number and age of diagnosis of the patient were previously reported in several tumor types and bulk blood in cancer patients[12]. The depth of sequencing did not influence our results, as WGS samples sequenced at 15X and 30X had comparable copy numbers (Fig. S2). However, the mitochondrial copy numbers did differ between the various cell types (Fig. 2). HSPCs displayed a mean mtDNA copy number of 481 (95% confidence interval: 396-566), whereas stem cells of the colon and intestine had a mean mtDNA copy number of 1213 (95% confidence interval: 1061-1367; $p < 0.0001$, linear mixed-effects model) and 958 (95% confidence interval: 856-1060; $p < 0.0001$, linear mixed-effects model), respectively. There was also a difference between colon and intestine ($p = 0.0013$). These differences likely reflect changes in mitochondrial activity between tissues[31,32]. The differences in mtDNA copy number between cell-types are consistent with contrasts found between various cancer types[12]. The high level of mtDNA copy numbers in intestinal stem cells is also consistent with the importance of mitochondria in these cells for proper stem cell functioning[33]. This observation indicates that the variation in mtDNA copy numbers between these cancers are not necessarily caused by mitochondrial dysfunction due to the malignant phenotype, but likely reflect the differences already found between healthy tissues from which these cancers arise.

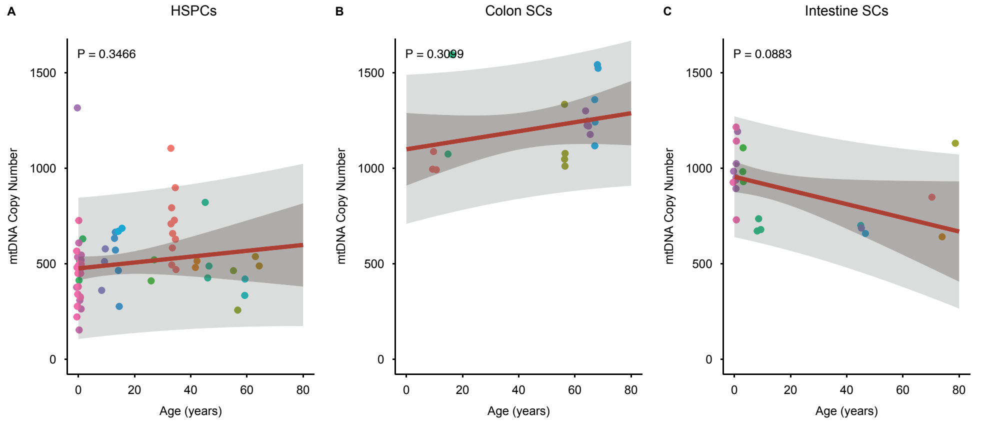


Figure 2 Effect of age and cell-type on mtDNA copy number.
a, **b**, and **c**The mtDNA copy number is plotted against the donor age for normal HSPCs (a) (62 samples; 14 donors), normal colon stem cells (b) (19 samples; 5 donors), and normal intestinal stem cells (c) (22 samples; 10 donors). *p*-values show the significance of the age of the donor on the mtDNA copy number (linear mixed-effects model). The red line indicates the mean fitted number of mutations at that age. The dark grey background shows the 95% confidence interval of the model, whereas the light grey background shows the 95% prediction interval. The prediction intervals show the predicted intervals that contain the mutation load of 95% of all cells in the population. A small amount of jitter was added to the dots to prevent them from completely overlapping. The color of the dots indicates the donor.

The mtDNA mutation burden in blood cancer is similar to normal stem cells

After investigating the mutation burden in mtDNA of normal cells, we compared how mutation accumulation was perturbed in mtDNA of cancers of the same tissue. First, we compared normal blood HSPCs to hematological cancers. We identified mutations in WGS data from our own lab and from samples of 15 patients from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) program that were sampled at diagnosis from patients with either acute myeloid leukemia or acute lymphoblastic leukemia[17,26,27]. Additionally, we analyzed the mtDNA mutation burden of patients with different hematological cancers (i.e., Lymph-BNHL, Lymph-CLL, Lymph-NOS, Myeloid-AML, Myeloid-MDS and Myeloid-MPN) whose data was included in the PCAWG consortium[12]. After combining this data, we found that the blood cancers had on average 0.5662 (95% confidence interval: 0.2933-0.8391) more base substitutions per sample than normal HSPCs from healthy donors after correcting for age ($p < 0.0001$; generalized linear model; Fig. 3a). This observation indicates that the mutation burden of blood cancer is only slightly higher than that of normal blood. However, this increased mutation burden was only present in a subset of cells. Most blood cancer samples harbored a similar number of mitochondrial substitutions as age-matched normal HSPCs, suggesting that the majority of mtDNA mutations in blood cancers are a consequence of normal age-related mutagenesis instead of a cancer-related mutator phenotype. Only one mutation in the blood cancers , which

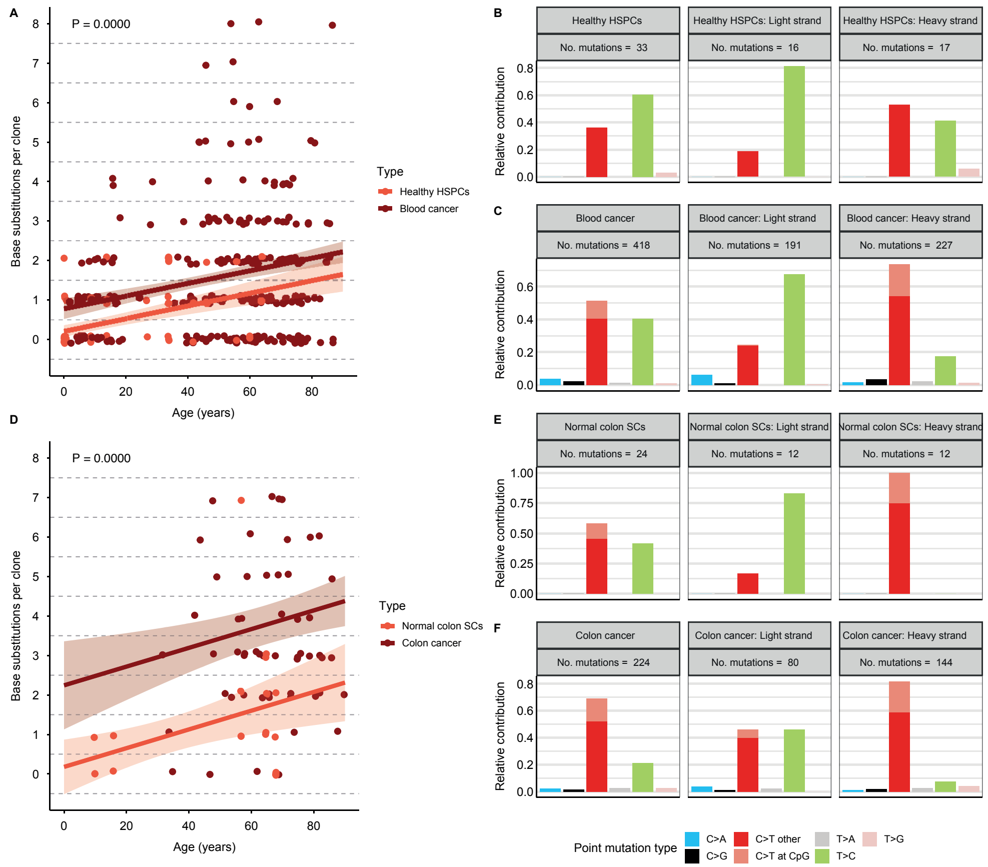


Figure 3 Comparison of mitochondrial substitutions between normal stem cells and cancers. **a** The number of mitochondrial base substitutions per clone is plotted against the donor age for normal HSPCs from healthy donors (33 mutations; 62 samples; 14 donors) and blood cancers (418 mutations; 264 samples; 264 donors). The *p*-value shows the significance of the difference in the number of substitutions between normal HSPCs and blood cancer (generalized linear model). The color of the dots and lines indicates the sample type. The trend lines indicate the mean fitted number of mutations at that age and sample type. The shaded backgrounds show the 95% confidence intervals of the model. A small amount of jitter was added to the dots to prevent them from completely overlapping. **b** and **c** 7-Spectrum of mitochondrial base substitutions for HSPCs from healthy donors (33 mutations; 62 samples; 14 donors) (b) and blood cancers (418 mutations; 264 samples; 264 donors) (c). A spectrum separated into light and heavy strands is also shown. The total number of base substitutions is indicated. **d** The number of mitochondrial base substitutions per clone is plotted against the donor age for normal colon stem cells (26 mutations; 19 samples; 5 donors) and colon cancers (224 mutations; 59 samples; 59 donors). The *p*-value shows the significance of the difference in the number of substitutions between normal colon and colon cancer (generalized linear model). The color of the dots and lines indicates the sample type. The trend lines indicate the mean fitted number of mutations at that age and sample type. The shaded backgrounds show the 95% confidence intervals of the model. A small amount of jitter was added to the dots to prevent them from completely overlapping. **e** and **f** 7-Spectrum of mitochondrial base substitutions for normal colon stem cells (24 mutations; 19 samples; 5 donors) (e) and colon cancers (224 mutations; 59 samples; 59 donors) (f). A spectrum separated into light and heavy strands is also shown. The total number of base substitutions is indicated.

was a synonymous base substitution, was present in both a normal HSPCs and blood cancer sample, indicating that the mutations we observed are random passengers and not recurrent drivers. The spectra of the mtDNA mutations in normal HSPCs consisted of mostly C>T substitutions on the heavy strand and T>C substitutions on the light strand (Fig. 3b). This spectrum is very similar to the spectrum of the PCAWG data, with a cosine similarity of 0.937 for the entire spectrum and cosine similarities of 0.984 and 0.920 for the light and heavy strands, respectively (Fig. 3c). Additionally, mutations in both the normal HSPCs and the blood cancers were distributed across the mitochondrial genome (Fig. S3a). These observations suggest that the mtDNA substitutions found in blood cancer samples are caused by the same mutational processes, likely related to mtDNA replication, as the substitutions found in normal blood, which further underlines the idea that most mtDNA mutations in blood cancer are the result of normal age-related mutagenesis[34]. Interestingly, we observed a slightly higher ratio of missense substitutions in the normal HSPCs compared to the blood cancers ($p = 0.048$; $X^2 = 4.7$; Chi-squared test; Fig. S3b). However, since this effect is small, it could be caused by differences in mutation calling or random chance.

The blood cancer samples with an elevated mutation load, here defined as samples with 4 or more substitutions, did not show an enrichment for any specific histological subtype ($p = 0.0545$; $X^2 = 13.8$; Chi-squared test). The mutation pattern of blood cancers with an increased mutation load was very similar to blood cancers with a lower mutation load, with a cosine similarity of 0.999 for the entire spectrum and cosine similarities of 0.997 and 0.999 for the light and heavy strands, respectively. The ratio of missense mutations was also similar between cancers with a higher and lower mutation load ($p = 0.311$; $X^2 = 1.2$; Chi-squared test), as was the distribution of mutations across the mitochondrial genome (Fig. S3c). These observations suggest that the increased mutation load found in some blood cancers is caused by an increased activity of the normal mutational processes found in mitochondria and not by a cancer specific mutational process.

Colon cancer has an increased mtDNA mutation burden

To test if these results generalize to more types of cancer, we compared colon cancers to normal colon stem cells. After correcting for age, colon cancers had on average 2.0661 (95% confidence interval: 1.2498-2.8824) more base substitutions per clone than normal colon ($P < 0.0001$; generalized linear model; Fig. 3d). The larger mean difference between normal and cancer samples in colon compared to blood was likely caused by a larger fraction of cancer samples having an elevated mutation load. The increased mitochondrial mutation load in cancer thus seems to be cancer type specific. There were no mutations that were present in both the normal colon stem cells and the colon cancers. The mutational spectra found in colon cancer showed an increased contribution of C>T mutations on the light strand, resulting in a decreased cosine similarity with normal colon (Fig. 3e, f; cosine similarity: 0.793). In contrast, the heavy strand had a high cosine similarity of 0.995. Since the normal colon samples did not contain many mutations, we also compared the mutation spectrum of the light strand of the colon cancer samples with that of the blood cancer samples. These spectra had a cosine similarity of 0.8970 and were significantly different from each other with $p = 0.0005$ ($X^2 = 23.477$; Chi-squared test). However, the trinucleotide profiles of the C>T mutations on the light strand are quite similar (Fig. S3d). This suggests that these C>T mutations are caused by the same process, which is, however, more active on the light strand of colon cancer samples compared to normal colon stem cells and blood cancers[34]. In line with this, the ratio of missense mutations was similar between colon cancer and normal colon stem cells ($p = 0.518$; $X^2 = 0.6$; Chi-squared test; Fig. S3e), and mutations in both groups were distributed across the mitochondrial genome (Fig. S3f).

The colon cancers with a mutation load of at least 4 substitutions had a similar mutation pattern as those with a lower mutation load with a cosine similarity of 0.929 on the light strand and 0.998 on the heavy strand. The ratio of missense mutations was also similar ($p = 0.464$; $X^2 = 0.6$; Chi-squared test) as was the distribution of mutations across the mitochondrial genome (Fig. S3g). This further supports our conclusion that the increased mutation load found in some cancers is the result of the normal mutational processes found in mitochondria.

Both blood and colon cancer had a lower mtDNA copy number than the corresponding normal stem cells ($p < 0.0000$; $p = 0.0001$; linear mixed-effects model; Fig. S3h, i). However, this is caused by technical differences in sequencing or sample preparation, since in-house pediatric AML samples had higher copy-numbers than pediatric AML samples from TARGET ($p < 0.0000$; $W = 149$; Wilcoxon rank sum test; Fig. S3j). Therefore, subsequent copy number analyses only included samples from our own lab.

Normal HSPCs of patients with cancer show an increased mutation accumulation

Since pediatric cancers are often characterized by an elevated mutation burden, which is caused by the presence of a mutational signature associated with oxidative stress[17], we hypothesized that the normal HSPCs of children with cancer could also have an increased mutation load in the mitochondria. To maximize our statistical power, we pooled together normal HSPCs at diagnosis, follow-up during remission, and the diagnosis of a secondary cancer, as we did not observe any differences between them (Methods). After correcting for age, normal HSPCs from children with leukemia had on average 0.3228 (95% confidence interval: 0.1429-0.5027) more substitutions per clone ($p = 0.0004$; generalized linear model; Fig. 4a). Similar to cancer cells, only a fraction of samples shows an increased mutation load. This observation was validated by an outlier test, which showed that the five samples out of 264 with 4 or more substitutions were all statistical outliers with $p<0.001$. The mutations in these samples had a median VAF of 0.0787, which is in a similar range as the median VAF of 0.0590 in HSPCs from leukemia patients with a lower mutational load ($p = 0.0719$; W=1610; Wilcoxon rank sum test). The presence of leukemic blasts in the bone marrow thus seems to result in an increased mutation load in the mitochondria of normal HSPCs.

Normal HSPCs from leukemia patients did not have a significantly different mtDNA copy number than HSPCs from healthy donors (Fig. S4a). However, one sample was a statistical outlier, with a mtDNA copy number of over 2000.

To further validate that the difference in mutation load between blood cancers and normal blood stem cells is small, we compared the mutation load of leukemias with the mutation load of normal HSPCs from the same patients. We did not observe a significant difference ($p = 0.5310$; generalized linear model; Fig. S4b); however, our statistical power was limited by the small number of patients for which both primary tumor samples and clonally expanded single-cell HSPCs were available.

Treatment does not result in an increased mutation load

The treatment of cancers can cause somatic mutations in the nuclei of normal cells, which has been associated with second primary cancers, which are cancers occurring in patients that have previously had a different primary cancer[16,27,35,36]. To investigate whether this also holds true for mtDNA, we analyzed mitochondrial mutations samples from children who received chemotherapy to treat pediatric cancer. The leukemia of patients with a secondary cancer did not contain an increased number of mitochondrial substitutions per clone compared to the normal HSPCs of healthy donors ($p = 0.6893$; generalized linear model; Fig. 4b). Secondary leukemias also did not contain an increased number of mitochondrial substitutions compared to primary leukemias from the same patient (Fig. S4c). One interesting hypothesis is that the lack of difference between HSPCs from healthy donors and patients with a secondary leukemia could be the result of damaged mitochondria having been cleared in patients with a second cancer[37].

Secondary leukemias did not have a significantly different mtDNA copy number than HSPCs from healthy donors (Fig. S4a). However, similar to the normal HSPCs, one

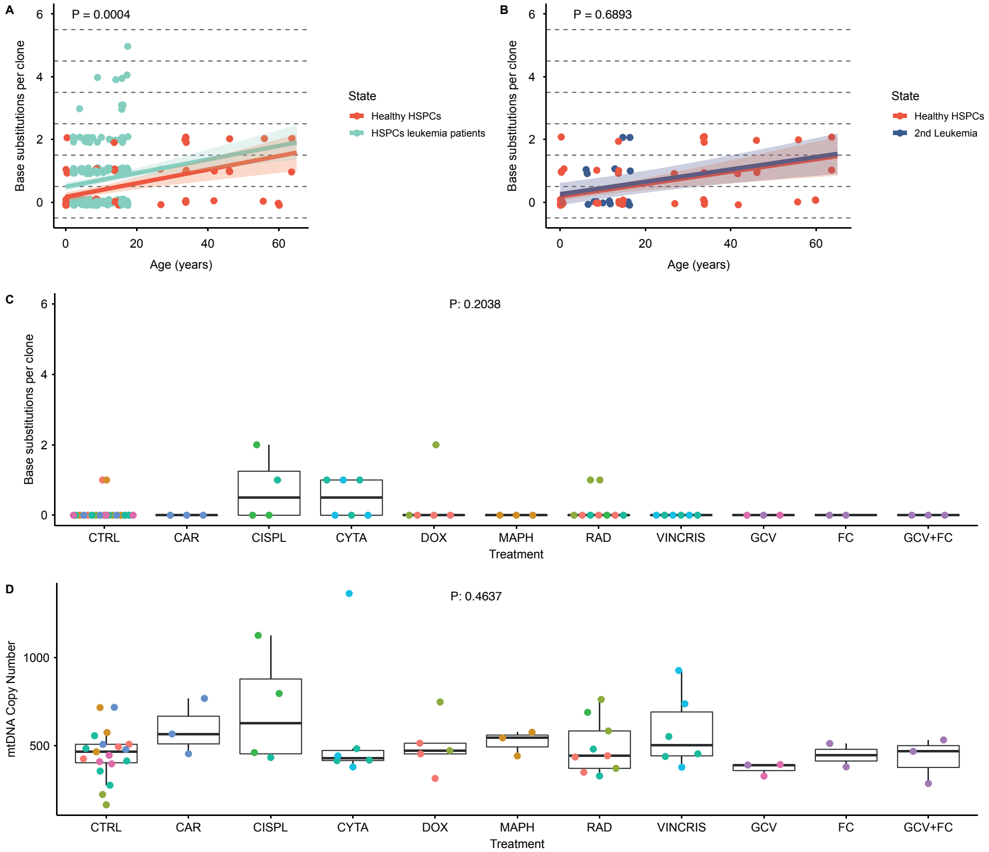


Figure 4 The effects of cancer and treatment on mitochondrial genomes.
a The number of mitochondrial base substitutions per clone is plotted against the donor age for HSPCs from healthy donors (33 mutations; 62 samples; 14 donors) and HSPCs at either diagnosis, follow-up during remission, or a diagnosis of a genetically unrelated secondary cancer (144 mutations; 202 samples; 28 donors). The p-values show the significance of the difference in the number of substitutions between HSPCs from healthy donors and HSPCs from leukemia patients (generalized linear model). The color of the dots and lines indicates the sample type. The trend lines indicate the mean fitted number of mutations at that age and sample type. The shaded backgrounds show the 95% confidence intervals of the model. A small amount of jitter was added to the dots to prevent them from completely overlapping. **b** The number of mitochondrial base substitutions per clone is plotted against the donor age for HSPCs from healthy donors (33 mutations; 62 samples; 14 donors) and secondary pediatric leukemias (8 mutations;16 samples; 16 donors). The p-values show the significance of the difference in the number of substitutions between HSPCs from healthy donors and secondary leukemias that are genetically unrelated from the original cancer (generalized linear model). A small amount of jitter was added to the dots to prevent them from completely overlapping. **c** The number of mitochondrial base substitutions per clone is shown for clonally expanded cord blood cells from healthy donors treated with different chemotherapies and X-ray. The color of the dots indicates the donor. CTRL = Control (2 mutations; 19 samples; 6 donors), CAR = carboplatin (0 mutations; 3 samples; 1 donor), CIS = cisplatin (3 mutations; 4 samples; 2 donors), CYTA = Cytarabine (3 mutations; 6 samples; 2 donors), DOX = Doxorubucin (2 mutations; 5 samples; 2 donors), MAPH = Maphosphamide (0 mutations; 3 samples; 1 donor), RAD = X-ray (2 mutations; 9 samples; 4 donors), VINCRIS = Vincristine (0 mutations; 6 samples; 2 donors), GCV = Ganciclovir (0 mutations; 3 samples; 2 donors), FC = Foscarnet (0 mutations; 2 samples; 1 donor), GCV+FC (0 mutations; 3 samples; 1 donor). **d** Comparison of the mtDNA copy numbers between clonally expanded cord blood cells from healthy donors treated with different chemotherapies and X-ray. The color of the dots indicates the donor. CTRL = Control (19 samples; 6 donors), CAR = carboplatin (3 samples; 1 donor), CIS = cisplatin (4 samples; 2 donors), CYTA = Cytarabine (6 samples; 2 donors), DOX = Doxorubucin (5 samples; 2 donors), MAPH = Maphosphamide (3 samples; 1 donor), RAD = X-ray (9 samples; 4 donors), VINCRIS = Vincristine (6 samples; 2 donors), GCV = Ganciclovir (3 samples; 2 donors), FC = Foscarnet (2 samples; 1 donor), GCV+FC (3 samples; 1 donor).

sample was a statistical outlier with a mtDNA copy number of over 2000. Interestingly, these outlier samples did not have high mutation loads.

To validate that treatment does not result in an increased mutation load in mitochondria we analyzed the WGS data of single CD34+ cord blood cells from healthy donors that were treated with chemotherapy, antiviral drugs, or X-ray *in vitro* for 3 days, after which they were clonally expanded[25,27,38]. While some of these treatments resulted in an increased mutation load in the nucleus, this was not the case for the mitochondrial genomes (*p* =0.2038; one-way ANOVA; Fig. 4c). This could be because mitochondrial genomes are not damaged by these treatments, damaged mitochondria are cleared, or mutations caused by treatment have a heteroplasmy level that is below the detection limit. Similar to the mutation load, we observed no differences in mtDNA copy numbers between samples that had been treated with different chemotherapies, antiviral drugs, or X-ray (*p* = 0.4637; one-way ANOVA; Fig. 4d).

Discussion

Here, we investigated the speed with which mitochondria in normal tissues accumulate somatic mutations with age and found that this was similar between different tissues. By comparing normal cells with cancer from the same tissue, we have also shown that most mitochondrial mutations in cancer are the result of normal mutagenesis and that treatment perturbations do not strongly impact the mitochondrial mutation load.

In general, cancers and treatment did not have a large effect on the mitochondrial genomes. Chemotherapy, for example, did not result in large observable increases in mitochondrial mutation loads both *in vivo* and *in vitro*, even though it can lead to large increases in nuclear mutation loads[25,27,35,39]. This suggests that the relation between cancer and mitochondria is not dependent on damage to the mitochondrial genome[3,12]. One possible explanation for the limited effect of cancer and its treatments is that the mitochondrial DNA damage they cause, might be resolved by cells clearing their damaged mitochondria. The increased mutation loads in pediatric cancer patients was present in only a subset of cells. This observation would not have been possible with bulk data and shows the advancement provided by single cell data.

Overall, our data suggests that damage to the mitochondrial genome is not a major driver for carcinogenesis.

Limitations of the study
Our approach for detecting mitochondrial variants could run into two issues. First, it can be difficult to distinguish *in vitro* and *in vivo* somatic mutations. Normally, this distinction is based on clonality, however mitochondrial somatic mutations that were present in the original single cell are unlikely to be clonal. In practice, these *in vitro* variants are unlikely to be an issue, because most of them are expected to have a very low VAF. This low VAF is the result of the high mtDNA copy number per clone and the lack of time for genetic drift to increase the VAF of these variants. Since we filter out all variants with a VAF below 0.1, most *in vitro* variants are likely removed. A consequence of this filtering is that we have likely missed some mutations with a very low level of heteroplasmy and underestimated the real mitochondrial mutation load. However, the low heteroplasmy level of these variants also makes it unlikely that they have a real biological effect[8].

A second issue is that selection or genetic drift can cause an inherited heteroplasmic variant to be lost or for its VAF to go below the detection limit in some or most cells. We have attempted to alleviate this issue by removing all variants for which there was any evidence in a matching bulk tissue, which should remove most germline variants. Additionally, since we filter out variants with a VAF below 0.01, small changes in the VAF of a heteroplasmic variant are insufficient to make it pass all filtering criteria in one sample, while being entirely undetectable in another sample. However, even with these controls, it is still possible that some somatic mutations were actually inherited, because somatic mutations are impossible to distinguish from heteroplasmic inherited variants with absolute certainty.

Overall, our study provides insights in the mitochondrial mutation accumulation and mtDNA copy numbers of normal cells and how this is perturbed by both cancer and treatment. These findings may contribute to our understanding of mitochondrial genomes and their role in disease.

STAR Methods

Lead contact
Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Ruben van Boxtel (R.vanBoxtel@prinsesmaximacentrum.nl).

Materials availability
This study did not generate new unique reagents.

Data availability
Data are available on EGA under accession numbers EGAS00001001682, EGAS00001000881, EGAS00001003068, EGAS00001003982, EGAS00001004593, EGAS00001004926, EGAS00001005141. The TARGET data are available on the database of Genotypes and Phenotypes (dbGaP) with accession number phs000218. The PCAWG mutation frequencies were provided by Young Seok.

Code availability
The NF-IAP can be found at https://github.com/UMCUGenetics/IAP. All original code can be found at: https://github.com/ProjectsVanBox/mitochondria_mutation_accumulation
Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Sample information
All 532 samples used in this study are available on EGA as described in the Data availability statement. The samples consist of clonally expanded single-cells and bulk samples that were sequenced using whole genome sequencing. In brief, the 19 normal colon, 22 healthy intestine, and 6 normal liver samples were clonally expanded from adult stem cells using organoid cultures[14]. Mitochondrial mutation accumulation was not investigated for the liver samples, because of a lack of statistical power. The 62 clonally expanded HSPC samples in the healthy blood group were generated by multiple studies. The samples from donors AC41, AC63, ACC55, AC33, BCH, and CB112 were generated using adult donors and cord blood[15]. The samples from donors MH2, NR1, and NR2 were generated from fetal blood[24]. The samples from SIB1, SIB2, SIB3, HAP1, and HAP2 were generated from hematopoietic stem cell transplant donors[25]. The 107 normal HSPCs in the diagnosis group and the 31 primary leukemia samples were generated from pediatric cancer donors. The samples from donors UPN025 to UPN033 were from AML patients[17], whereas the samples from UPN001 to UPN023 came from patients with a variety of primary cancers[27]. The samples from the donors PAMXZY, PAMYMA, PANZLR, PARBTV, PARXYR, PASDKZ, PASDXR, PASFHK, PASFJJ, PASLZE, PASSLT, PASVJS, PASYWA, PATISD, and PATKKJ were from TARGET[26]. The 36 normal HSPCs in the follow-up group, the 59 normal HSPCs in the diagnosis 2 group, and the 16 secondary leukemias were sampled at either the time of primary cancer remission, or during the diagnosis of a second cancer

that is genetically distinct from the original[27]. The 36 HSPC samples in the hematopoietic stem cell transplant (HSCT) recipient group were generated from donors that had received a hematopoietic stem cell transplant[25]. Since these HSPCs were mostly extracted from peripheral blood instead of bone marrow, they could not be directly compared to the HSPC samples in the healthy blood group. However, their inclusion could still aid in the filtering of somatic mutations because any variants present in both these HSPCs and the HSPCs of matching donors is likely to be an inherited variant and not a somatic mutation. The 63 HSPC samples in the *in vitro* chemotherapy group were generated from cord blood[25,27,38]. Bulk cord blood was treated with the relevant treatment for 3 days, after which a single HSPC was clonally expanded and sequenced[38]. The samples that we analyzed with the mitochondrial pipeline were compared to previously identified substitutions in adult cancers from PCAWG[12].

Read alignment and variant calling

Some samples were originally aligned to hg19. These samples were re-aligned to hg38 by first converting them to the FASTQ format using Picard SamToFastq with the “RG_TAD=ID” and “OUTPUT_PER_RG=true” arguments (v2.24.1)[40]. The samples were then compressed using bgzip (v1.0)[41]. Finally, read alignment was performed using the Nextflow Illumina Analysis Pipeline (NF-IAP, v1.2).

The bulk skin biopsies of N01, NR1 and NR2 were sequenced on both Illumina HiSeq X Ten sequencers and Nova sequencers. The resulting BAM files were subsequently merged using samtools merge (v1.3) and the library (LB) and sample (SM) fields were unified for each readgroup in the new bamfile, as previously described[24]. After converting these samples to hg38, this merging step was repeated.

Samples were analyzed using a modified version of the Broad Institute’s GATK (v4.1.3.0) Mitochondria pipeline (https://github.com/broadinstitute/gatk/tree/master/scripts/mitochondria_m2_wdl)[28]. This pipeline is written in WDL and was run on a high-performance compute facility using the Cromwell execution engine. In short, this pipeline takes hg38 BAM files as its input and subsets them to only the mitochondrial reads, which includes removing reads mapping to the NuMTs[42]. The pipeline also takes mean nuclear coverage as its input, which was calculated by the NF-IAP using GATKs CollectWGSMetrics tool. The mitochondrial reads are then aligned twice to both the regular and a shifted version of the mitochondrial genome to overcome the problem of linear mapping to a circular genome. After this, the haplochecker command from mitolib (v0.1.2; https://github.com/haansi/mitolib) is used to identify the haplotype of a sample and detect any contamination. Next, variants are called for both alignments using the mitochondria mode of Mutect2. The variants called on the shifted version of the mitochondrial genome are then lifted over using Picard LiftoverVcf (v2.20.1) and merged with the variants called on the regular mitochondrial genome using Picard MergeVcfs (v2.20.1). The “.stats” files from Mutect2 are merged using GATKs MergeMutectStats. Next, artifacts are flagged using GATKs FilterMutectCalls using the “mitochondria-mode” argument, which includes the “ChimericOriginalAlignmentFilter” and “PolymorphicNuMTFilter” filters. Additionally, the “autosomal-coverage” argument was supplied with the mean nuclear coverage, the “stats” argument was supplied with the merged “.stats” file, and the “contatmination-estimate” argument was supplied with the estimated contamination from mitolib. This step also ensures that false positive variants caused by NuMTs are flagged by using a Poisson distribution based on the mean nuclear coverage[29,43,44]. We modified the pipeline to also flag common variants using haplotype specific blacklists from MITOMAP (v102) with the “blacklisted_site” flag using GATKs VariantFiltration[45]. Next, the NF-IAP was used to annotate the identified variants. These annotations include a prediction of the effect of the variants by SnpEff (v4.3t)[46].

Three samples (MH2LIMPPCL13, PMC21636MPP6, and PAKIYWBMPC) were removed, because their mean mitochondrial read coverage was lower than a thousand, indicating potential technical issues and making it more likely for low VAF variants to be missed. As a result, there was only a bulk sample (PAKIYWBMNF) left for one donor, which was therefore also removed.

Somatic variant filtering

The R language (v3.6.3) was used to filter for somatic variants and also perform all subsequent analyses[47]. We only considered heterozygous variants that had passed all quality filters of the mitochondria GATK pipeline and did not have multiple alternative alleles. This meant that variants with any of the following filtering flags were removed: “blacklisted_site”, “strand_bias”, “base_qual”, “weak_evidence”, “numt_novel”, “position”, and “contamination”. Next, we compared the genotypes of these variants across all samples. Variants that were present in a subset, but not all samples of a donor were considered to be somatic. Furthermore, variants that were called in a sample and a matching bulk sample were removed to prevent inherited heteroplasmic variants from being incorrectly identified as somatic mutations. Since the cord blood samples treated with either chemotherapy, anti-viral drugs, or X-rays did not have matching bulk samples, any shared variants in them were removed. Next, variants present in more than one donor were filtered out, because they could be sequencing artifacts or recurrent false positives caused by NuMTs[12]. Variants that were previously filtered out were included in the preceding comparisons between samples as a control. This prevents germline variants and sequencing artifacts from being called as true somatic mutations when they were flagged in some samples. Additionally, variants with a VAF below 0.01 were removed. This step removes false positives caused by NuMTs, as they are expected to have a low VAF[42]. Next, any variants that were previously found to be likely false positives caused by NuMTs were removed[42,44]. For this step we used variants in the RHO94 database that were found in capture-enrichment data from Li et al., 2012 and variants that were likely false positives caused by more recent NuMT insertions from Dayama et al., 2014. Finally, variants that occurred in more than one sample were manually inspected. Variants in bulk samples were identified, but not used for subsequent analyses.

Mutation load accumulation

Since mutations can already accumulate before birth, the age of all samples was calculated from conception. A Poisson generalized linear model with an “identity” link function was fitted to the determine the effect of age on the number of base substitutions in normal clonal blood samples, using the following command: “glm(freq ~ age, data = healthy_freq, family = poisson(link = "identity"))”. A Poisson distribution was used, because mutation accumulation is expected to be a Poisson process, as mutations are discrete and generally independent events. Additionally, linear models are not well suited for the small and discrete numbers of mutations found in mitochondria. A mixed-effects Poisson model with an “identity” link function was also attempted, however this failed to converge. This model was called with the following command: “glm_mixed_log_m <- glmer(freq ~ age + (0 + age | patient), data = healthy_freq, family = poisson(link = "identity"))”. To validate our modeling choices, we also fitted several other models. These included, a zero-inflated Poisson model (command: “zeroinfl(freq ~ age, data = healthy_freq, dist="poisson")"), a linear model (command: “lm(freq ~ age, data = healthy_freq)”), a linear mixed-effects model with a random slope (command: " lme(freq ~ age, random = ~-1 + age | patient, data = healthy_freq)”) and three Poisson models with a “log” identity link (commands: “glmer(freq ~ age + (0 + age | patient), data = healthy_freq, family = poisson(link = "log"))”; “glmer(freq ~ age + (1 | patient), data = healthy_freq, family = poisson(link = "log"))”; “glm(freq ~ age, data = healthy_freq, family = poisson(link = "log"))”;). The first two of these Poisson models were mixed-effects models with respectively a random slope and a random intercept. Our main model was superior to these alternative models based on both the Bayesian information criterion and the Akaike information criterion. We would like to note that the age variable was significant not just in the main model, but also in all other models, except for the zero-inflated one. Models with the same form as the main normal blood model were fitted for normal colon and normal intestine samples.

To determine the effects of cell-type, treatment, and disease, relevant samples were compared with the normal clonal HSPC samples. For each analysis of cell-type, treatment, or disease a model was fitted that is similar to the base model, but with an extra explanatory variable for cell-type or for the treatment or disease of interest. The commands to calculate these models had the following structure: “glm(freq ~ cell-type/treatment/disease + age, data = data_set, family = poisson(link = "identity"))”. To compare the mutation loads of leukemia with HSPCs at diagnosis from the same patient we generated a model without the age variable, using the following command: “glm(freq ~ state_name + patient, family = poisson(link = "identity"), data = dx1_vs_leukemia_freq)”. Mixed-effects models were fitted using a combination of the nlme (v3.1-148) and lme4 (v1.1-23) R packages[48,49].

To calculate the confidence and prediction intervals of a Poisson model, a grid was made of possible input variables. For the confidence intervals, the fit and standard error (se) were then calculated on the scale of the linear predictors for each point in the grid, using the command: “predict(model, type = "link", se = T, newdata = x_grid)”. The confidence interval was then defined as the fit +- 1.96 * se. Since the “identity” link was used the confidence interval did not need to be converted back to the response variable scale. To calculate the prediction interval of a Poisson model the underlying data was bootstrapped 10,000 times, using the sample function with the “replace = TRUE” argument. In each bootstrap iteration the model is updated with the bootstrapped data using the update function and the fit is calculated for each point in the grid. These estimated values were then used as the lambda to generate a random number from a Poisson distribution for each point in the grid, using the “rpois” function. The generated numbers across all bootstrap iterations were then combined and the 2.5% and 97.5% quantiles were then used as the prediction interval.

To identify any differences in the mutational load of the HSPCs from patients at diagnosis compared to patients at follow-up during remission, or patients at the diagnosis of a secondary cancer, we fit two Poisson models using only these HSPCs, regressing the mutation load against the donors age as described above. In the second model we included a variable, describing the patient’s disease state at sampling. This model was inferior compared to the model without this variable based on the Bayesian Information Criterion and the Akaike Information Criterion.

Comparison mitochondrial and nuclear mutation load

The predicted number of mitochondrial mutations was calculated using the Poisson model trained on HSPCs from healthy donors. The predicted number of nuclear mitochondrial mutations was calculated using a linear mixed-effects model, using the following command: lme(norm_muts ~ age, random = ~ -1 + age | patient, data = nuclear_mito_muts). Samples were then classified as having either more or less mutations than predicted by these models. A Pearson's Chi-squared test was then performed to see if the mitochondrial and nuclear classifications were independent. Chi-squared tests were calculated using Monte Carlo simulations with 2000 replicates, using the chisq.test function with the "simulate.p.value" argument.

Copy number analysis

The number of mitochondrial genomes per clone was calculated by dividing the mean read coverage in the mitochondrial genome by the mean read coverage in the nuclear genome and multiplying by two. A linear mixed-effects model with a random slope was fitted to determine the effect of age on the mitochondrial copy number, using the following command: "lme(cnv_mean ~ age, random = ~ 0 + age | patient, data = healthy_cnv)". A model combining the blood, colon and intestine was fitted using: "lme(cnv_mean ~ age * state, random = ~ 0 + age | patient, data = tissue_cnv)". This model included an interaction between the cell type and age, to allow for differences in the slopes between the cell types. Because the age variable was not significant, it was not used in subsequent copy number models. To determine the effects of treatment and disease, relevant samples were compared with the healthy clonal samples. For each analysis of a treatment or disease a linear mixed-effects model was fitted with the treatment or disease included as an explanatory variable. The ggeffects (v0.15.0) package was used to calculate the confidence and prediction intervals of linear mixed-effects models[50].

Outliers in the models were detected by calculating the odds of the standardized absolute residuals occurring under a standard normal distribution for both the mutation load and copy number models. The command for this was: "multiply_by(pnorm(-multiply_by(abs(resid(model, type = "pearson")), -1)), 2)".

Mutation spectra

Mutation spectra and their cosine similarities were calculated and visualized using the MutationalPatterns (v3.3.4) R package[51]. Spectra were calculated separately for substitutions with a "C" or "T" reference base and substitutions with a "G" or "A" reference base. The "type_context" function from MutationalPatterns was used to identify for each C>T mutation, whether it occurred within a CpG context. Variants that were shared between multiple samples of a single donor were only counted once.

Missense mutations were identified based on SnpEff annotations. "start_lost", "stop_gained", and "stop_lost" mutations were also considered as missense mutations. For the variants identified by PCAWG, the supplied PCAWG annotations were used. The ratios of missense and other mutations, consisting of both synonymous and non-coding variants, were compared between groups using Pearson's Chi-squared tests as described above.

Comparisons to PCAWG

To determine the effect of adult cancer on mitochondrial mutation accumulation, normal clonally expanded HSPCs from healthy donors were compared to blood cancer mutations from the PCAWG consortium. Lymph-BNHL, Lymph-CLL, Lymph-NOS, Myeloid-AML, Myeloid-MDS, and Myeloid-MPN cancers samples were pooled together in a single blood cancer category. Normal clonally expanded colon stem cells were compared to mutations in ColoRect-AdenoCA cancer samples, which were referred to as colon cancer. Mutation accumulation, copy numbers and mutation spectra were analyzed as described above. Comparisons were also made between blood and colon cancer samples with at least 4 substitutions and cancer samples with a lower mutation load.

Gene mutations

The number of mutations per gene was calculated per sample category. For each category only clonal samples were included. The mutation counts were normalized by dividing the number of mutations per gene by the gene lengths, which, together with the gene strands, were obtained via Ensembl (v104)[52].

Quantification and statistical analysis

The numbers of samples and donors per analysis are indicated in the figure legends. *p*-values are indicated in the figures and explained in the figure legends and main text. To assess the significance of mitochondrial mutation accumulation, generalized linear models with a Poisson distribution and a "identity" link function were used. Linear mixed-effects models were used to assess the significance of mtDNA copy number differences between groups and to assess the effect of age on this variable. Cosine similarities were used to compare mutation spectra and read distributions. One-way ANOVAs were used to assess the significance of differences between cord-blood clones treated with different chemotherapies and X-ray. To assess if the nuclear and mitochondrial mutation loads were significantly related a Pearson's Chi-squared was used. A Chi-squared test was also used to compare the histology between blood cancers with 4 or more substitutions and blood cancers with a lower mutation load. Furthermore, Chi-squared tests were performed to compare the ratio of missense versus other mutations between groups. Chi-squared tests were calculated using Monte Carlo simulations with 2000 replicates. To assess the significance of the correlation between the maximum VAF of a donor and the donors' age a linear regression was used. To assess the significance of the difference between mtDNA copy numbers of in-house samples and samples from TARGET a Wilcoxon rank sum test with continuity correction was used. A Wilcoxon rank sum test with continuity correction was also used to compare the VAFs of mutations in HSPC samples with a high mutation load of 4 or more substitutions with HSPC samples with a lower mutation load.

Acknowledgements

We would like to thank Jayne Y. Hehir-Kwa for her help in setting up the mitochondrial pipeline. Furthermore, we want to thank Arne van Hoeck for sharing the bam files of the normal colon and intestine cells. We also want to thank Young Seok for sharing a list of all PCAWG samples for which mitochondrial mutations were explored. Finally, we want to thank Axel R. Huber, Jurrian de Kanter, Flavia Peci, Arianne M. Brandsma, Rurika Oka and Markus J. van Roosmalen for their help in ordering and interpreting the data. This study was financially supported by a VIDI grant from the Dutch Research Council (NWO no. 016.Vidi.171.023) to R.B.

Author contributions

F.M. and J.T.D. gathered the data and performed bioinformatic analyses. F.M. and R.B. wrote the manuscript. R.B. designed and supervised the study.

Declaration of interests

Authors declare no competing interests.

References

1. Hengartner MO. The biochemistry of apoptosis. Nature. 2000;407(6805):770–6.
2. Duchen MR. Mitochondria and calcium: from cell signalling to cell death. J Physiol. 2000 Nov 15;529 Pt 1(Pt 1):57–68.
3. Zong W-X, Rabinowitz JD, White E. Mitochondria and Cancer. Mol Cell. 2016 Mar 3;61(5):667–76.
4. Schon EA, DiMauro S, Hirano M. Human mitochondrial DNA: roles of inherited and somatic mutations. Nat Rev Genet. 2012 Dec;13(12):878–90.
5. García-Rodríguez LJBT-M in CB. Appendix 1. Basic Properties of Mitochondria. In: Mitochondria, 2nd Edition. Academic Press; 2007. p. 809–12.
6. Satoh M, Kuroiwa T. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Exp Cell Res. 1991;196(1):137–40.
7. Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290(5806):457–65.

8. Stewart JB, Chinnery PF. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat Rev Genet. 2015;16(9):530–42.
9. Alston CL, Rocha MC, Lax NZ, Turnbull DM, Taylor RW. The genetics and pathology of mitochondrial disease. J Pathol. 2016/11/02. 2017 Jan;241(2):236–50.
10. Greaves LC, Reeve AK, Taylor RW, Turnbull DM. Mitochondrial DNA and disease. J Pathol. 2012 Jan 1;226(2):274–86.
11. Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. Nat Rev Genet. 2005;6(5):389–402.
12. Yuan Y, Ju YS, Kim Y, Li J, Wang Y, Yoon CJ, et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nat Genet. 2020;52(3):342–52.
13. Smith AL, Whitehall JC, Bradshaw C, Gay D, Robertson F, Blain AP, et al. Age-associated mitochondrial DNA mutations cause metabolic remodelling that contributes to accelerated intestinal tumorigenesis. Nat cancer. 2020/09/21. 2020 Oct;1(10):976–89.
14. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016 Oct;538(7624):260–4.
15. Osorio FG, Rosendahl Huber A, Oka R, Verheul M, Patel SH, Hasaart K, et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. 2018 Nov;25(9):2308-2316.e4.
16. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;574(7779):532–7.
17. Brandsma AM, Bertrums EJM, van Roosmalen MJ, Hofman DA, Oka R, Verheul M, et al. Mutation Signatures of Pediatric Acute Myeloid Leukemia and Normal Blood Progenitors Associated with Differential Patient Outcomes. Blood Cancer Discov. 2021 Sep 1;2(5):484 LP – 499.
18. Williams SL, Mash DC, Züchner S, Moraes CT. Somatic mtDNA Mutation Spectra in the Aging Human Putamen. PLOS Genet. 2013 Dec 5;9(12):e1003990.
19. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet. 2013/09/26. 2013;9(9):e1003794–e1003794.
20. Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M, et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife. 2014 Oct 1;3:e02935.
21. Baker KT, Nachmanson D, Kumar S, Emond MJ, Ussakli C, Brentnall TA, et al. Mitochondrial DNA Mutations are Associated with Ulcerative Colitis Preneoplasia but Tend to be Negatively Selected in Cancer. Mol Cancer Res. 2018/11/16. 2019 Feb;17(2):488–98.
22. Greaves LC, Nooteboom M, Elson JL, Tuppen HAL, Taylor GA, Commane DM, et al. Clonal Expansion of Early to Mid-Life Mitochondrial DNA Point Mutations Drives Mitochondrial Dysfunction during Human Ageing. PLOS Genet. 2014 Sep 18;10(9):e1004620.
23. Lawless C, Greaves L, Reeve AK, Turnbull DM, Vincent AE. The rise and rise of mitochondrial DNA mutations. Open Biol. 2020/05/20. 2020 May;10(5):200061.
24. Hasaart KAL, Manders F, van der Hoorn M-L, Verheul M, Poplonski T, Kuijk E, et al. Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis. Sci Rep. 2020;10(1):12991.
25. de Kanter JK, Peci F, Bertrums E, Rosendahl Huber A, van Leeuwen A, van Roosmalen MJ, et al. Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. Cell Stem Cell. 2021;
26. Bolouri H, Farrar JE, Triche Jr T, Ries RE, Lim EL, Alonzo TA, et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat Med. 2017/12/11. 2018 Jan;24(1):103–12.
27. Bertrums EJM, Rosendahl Huber AKM, de Kanter JK, Brandsma AM, van Leeuwen AJCN, Verheul M, et al. Elevated Mutational Age in Blood of Children Treated for Cancer Contributes to Therapy-Related Myeloid Neoplasms. Cancer Discov. 2022 Aug 5;12(8):1860–72.

28. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep 1;20(9):1297–303.
29. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. bioRxiv. 2019;861054.
30. Mourier T, Hansen AJ, Willerslev E, Arctander P. The Human Genome Project Reveals a Continuous Transfer of Large Mitochondrial Fragments to the Nucleus. Mol Biol Evol. 2001 Sep 1;18(9):1833–7.
31. Fernández-Vizarra E, Enríquez JA, Pérez-Martos A, Montoya J, Fernández-Silva P. Tissue-specific differences in mitochondrial activity and biogenesis. Mitochondrion. 2011;11(1):207–13.
32. Herbers E, Kekäläinen NJ, Hangas A, Pohjoismäki JL, Goffart S. Tissue specific differences in mitochondrial DNA maintenance and expression. Mitochondrion. 2019;44:85–92.
33. Rodríguez-Colman MJ, Schewe M, Meerlo M, Stigter E, Gerrits J, Pras-Raves M, et al. Interplay between metabolic identities in the intestinal crypt supports stem cell function. Nature. 2017;543(7645):424–7.
34. Sanchez-Contreras M, Sweetwyne MT, Kohrn BF, Tsantilas KA, Hipp MJ, Schmidt EK, et al. A replication-linked mutational gradient drives somatic mutation accumulation and influences germline polymorphisms and genome composition in mitochondrial DNA. Nucleic Acids Res. 2021 Nov 8;49(19):11103–18.
35. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177:821-836.e16.
36. Christensen S, Van der Roest B, Besselink N, Janssen R, Boymans S, Martens JWM, et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun. 2019;10(1):4571.
37. Zong W-X, Rabinowitz JD, White E. Mitochondria and Cancer. Mol Cell. 2016 Mar 3;61(5)
38. Rosendahl Huber A, van Leeuwen AJCN, Peci F, de Kanter JK, Bertrums EJM, van Boxtel R. Whole-genome sequencing and mutational analysis of human cord-blood derived stem and progenitor cells. STAR Protoc. 2022;3(2):101361.
39. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
40. Picard toolkit. Broad Institute, GitHub repository. Broad Institute; 2019.
41. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience. 2021 Feb 16;10(2):giab007.
42. Li M, Schroeder R, Ko A, Stoneking M. Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs. Nucleic Acids Res. 2012 Oct 1;40(18):e137–e137.
43. Santibanez-Koref M, Griffin H, Turnbull DM, Chinnery PF, Herbert M, Hudson G. Assessing mitochondrial heteroplasmy using next generation sequencing: A note of caution. Mitochondrion. 2018/08/09. 2019 May;46:302–6.
44. Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res. 2014/10/27. 2014 Nov 10;42(20):12640–9.
45. Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, et al. mtDNA Variation and Analysis Using Mitomap and Mitomaster. Curr Protoc Bioinforma. 2013 Dec;44(123):1.23.1-1.23.26.
46. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.
47. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2018.
48. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. {nlme}: Linear and Nonlinear Mixed Effects Models. 2018.
49. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Software; Vol 1, Issue 1. 2015;
50. Lüdecke D. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. J Open Source Softw. 2018;3(26):772.
51. Manders F, Brandsma AM, de Kanter J, Verheul M, Oka R, van Roosmalen MJ, et al. MutationalPatterns: the one stop shop for the analysis of mutational processes. BMC Genomics. 2022;23(1):134.
52. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;48:D682–8.

Supplementary material

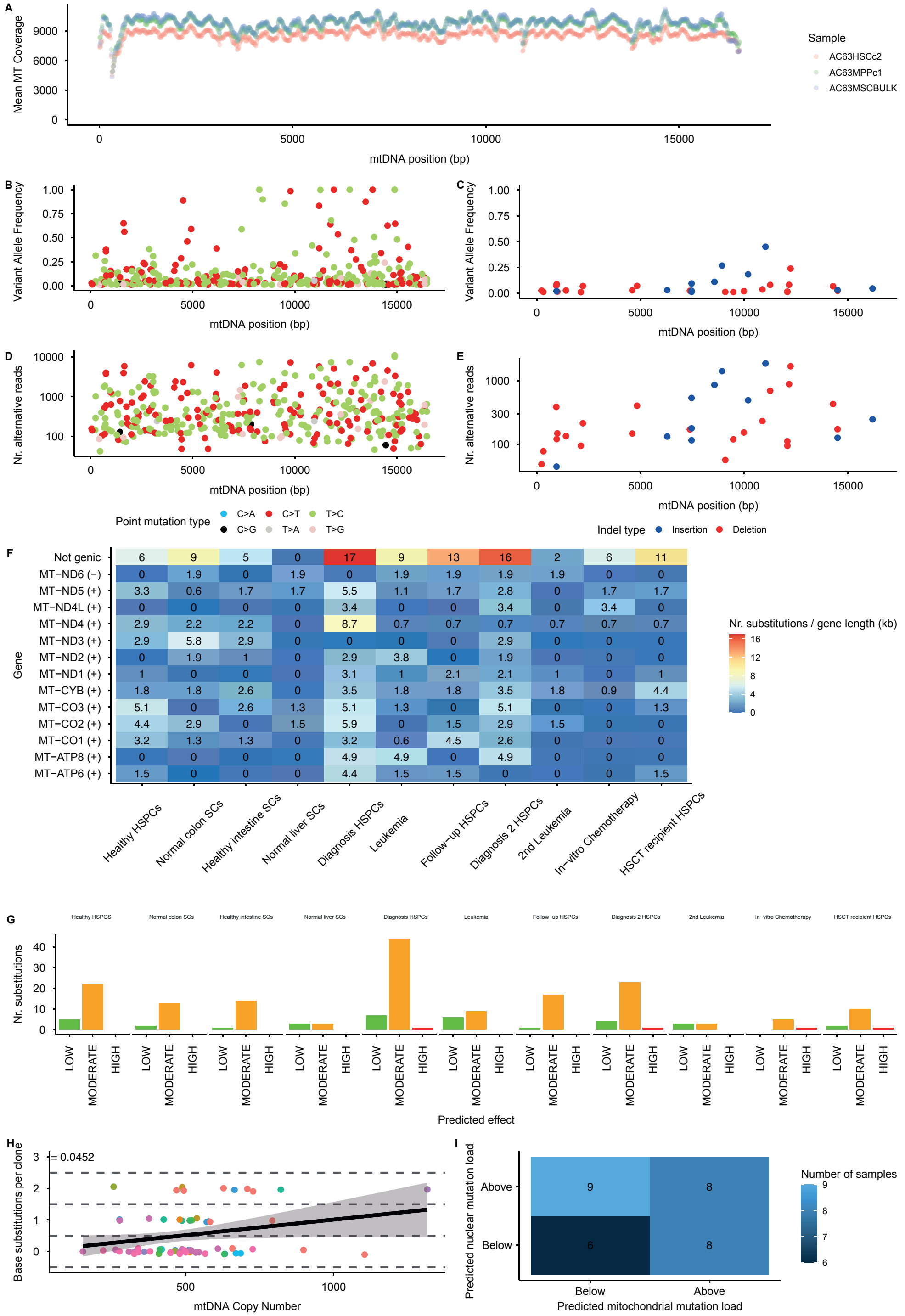


Figure S1. Detection of mitochondrial variants. Related to Figure 1.

a The mean mitochondrial read coverage is plotted in 20bp windows across the genome for 3 samples of donor AC63. Each dot represents a single 20bp window. The colors of the dots indicate the samples. **b** The variant allele frequency of substitutions from all samples (370 mutations; 532 samples; 88 donors) is plotted against their position on the mitochondrial genome. The color of the dots indicates the base substitution type. **c**The variant allele frequency of indels from all samples (33 mutations; 532 samples; 88 donors) is plotted against their position on the mitochondrial genome. The color of the dots indicates whether a variant is an insertion or a deletion. **d** The number of alternative reads of substitutions from all samples (370 mutations; 532 samples; 88 donors) is plotted against their position on the mitochondrial genome. The color of the dots indicates the base substitution type. **e** The number of alternative reads of indels from all samples (33 mutations; 532 samples; 88 donors) is plotted against their position on the mitochondrial genome. The color of the dots indicates whether a variant is an insertion or a deletion. **f** Heatmap depicting the number of substitutions divided by the gene length in kilobases per gene and for each of the following sample groups: Healthy blood (33 mutations; 62 samples; 14 donors), Normal colon (24 mutations; 19 samples; 5 donors), Healthy intestine (20 mutations; 22 samples; 10 donors), Normal liver, (6 mutations; 6 samples; 2 donors), Diagnosis (67 mutations; 107 samples; 17 donors), Leukemia (24 mutations; 31 samples; 31 donors), Follow-up (31 mutations; 36 samples; 8 donors), Diagnosis 2 (44 mutations; 59 samples; 11 donors), *In vitro* Chemotherapy (8 mutations; 63 samples; 9 donors), HSCT recipient (23 mutations; 36 samples; 9 donors). Substitutions within a group are pooled. Genes with a "+" behind their name are located on the light strand, whereas genes with a "-" are located on the heavy strand. **g** The effect of genic substitutions predicted by SnpEff for each of the following sample groups: Healthy blood (33 mutations; 62 samples; 14 donors), Normal colon (24 mutations; 19 samples; 5 donors), Healthy intestine (20 mutations; 22 samples; 10 donors), Normal liver, (6 mutations; 6 samples; 2 donors), Diagnosis (67 mutations; 107 samples; 17 donors), Leukemia (24 mutations; 31 samples; 31 donors), Follow-up (31 mutations; 36 samples; 8 donors), Diagnosis 2 (44 mutations; 59 samples; 11 donors), *In vitro* Chemotherapy (8 mutations; 63 samples; 9 donors), HSCT recipient (23 mutations; 36 samples; 9 donors). **h** The number of base substitutions per clone is plotted against the mtDNA copy number for HSPCs from healthy donors (33 mutations; 62 samples; 14 donors). A small amount of jitter was added to the dots to prevent them from completely overlapping. The color of the dots indicates the donor. The *p*-value shows the significance of the mtDNA copy number on the number of substitutions per clone (generalized linear model). **i** Heatmap depicting the number of HSPCs from healthy donors (samples = 31; donors = 9) whose mutation load is above or below the predicted mutation load in the mitochondrial and nuclear genomes. The predictions are based on regression models. Only samples for which both the mitochondrial and nuclear mutation loads had been determined were used.

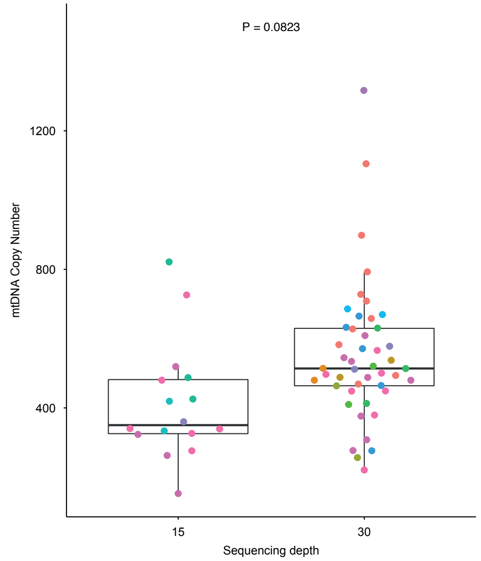


Figure S2. mtDNA copy numbers in normal clones. Related to Figure 2.
Comparison of the mtDNA copy numbers between samples that were sequenced at 15x (20 samples; 5 donors) or 30x (53 samples; 12 donors). The *p*-value shows the significance of the sequencing depth on the mtDNA copy number (linear mixed-effects model).

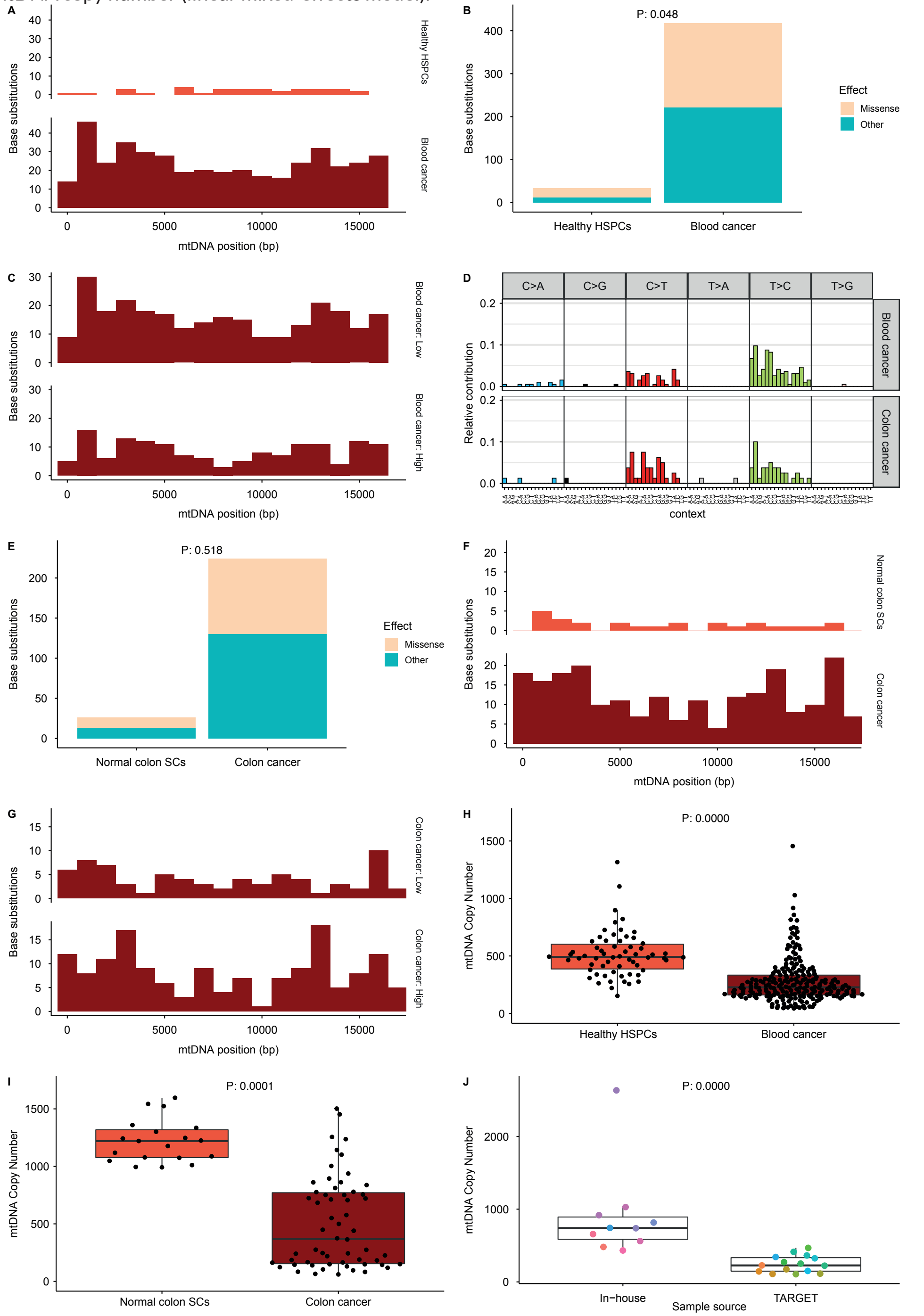


Figure S3. Mutation profiles and difference in mtDNA copy number between normal and cancer cells. Related to Figure 3.
a The total number of base substitutions across the genome is shown in 1kb bins for both HSPCs from healthy donors (33 mutations; 62 samples; 14 donors) and blood cancers (418 mutations; 264 samples; 264 donors). **b** The total number of missense and other base substitutions is shown for both HSPCs from healthy donors (33 mutations; 62 samples; 14 donors) and blood cancers (418 mutations; 264 samples; 264 donors). The color indicates whether substitutions are missense or other (synonymous/non-coding). **c** The total number of base substitutions across the genome is shown in 1kb bins for both blood cancers with a high mutation load of at least 4 substitutions (148 mutations; 29 samples; 29 donors) and blood cancers with a lower mutation load (270 mutations; 235 samples; 235 donors). **d** 96-Profile of mitochondrial base substitutions on the light strand for blood cancers (194 mutations; 264 samples; 264 donors) and colon cancers (80 mutations; 59 samples; 59 donors). **e** The total number of missense and other base substitutions is shown for both normal colon stem cells (26 mutations; 19 samples; 5 donors) and colon cancers (224 mutations; 59 samples; 59 donors). The color indicates whether substitutions are missense or other (synonymous/non-coding). **f** The total number of base substitutions across the genome is shown in 1kb bins for both normal colon stem cells (26 mutations; 19 samples; 5 donors) and colon cancers (224 mutations; 59 samples; 59 donors). **g** The total number of base substitutions across the genome is shown in 1kb bins for both colon cancers with a high mutation load of 4 or more substitutions (151 mutations; 25 samples; 25 donors) and colon cancers with a lower mutation load (73 mutations; 34 samples; 34 donors). **h** Comparison of the mtDNA copy numbers between HSPCs from healthy donors (62 samples; 14 donors) and blood cancers (229 samples; 229 donors). **i** Comparison of the mtDNA copy numbers between normal colon stem cells (19 samples; 5 donors) and colon cancers (58 samples; 58 donors). The *p*-values show the significance of the difference between normal and cancer cells (linear mixed-effects model). A small amount of jitter was added to the dots to prevent them from completely overlapping. **j** Comparison of the mtDNA copy numbers between AMLs from our own lab (10 samples; 10 donors) and AMLs from TARGET (15 samples; 15 donors). The color of the dots indicates the donor. A small amount of jitter was added to the dots to prevent them from completely overlapping. The *p*-value shows the significance of the difference between the two groups (Wilcoxon rank sum test).

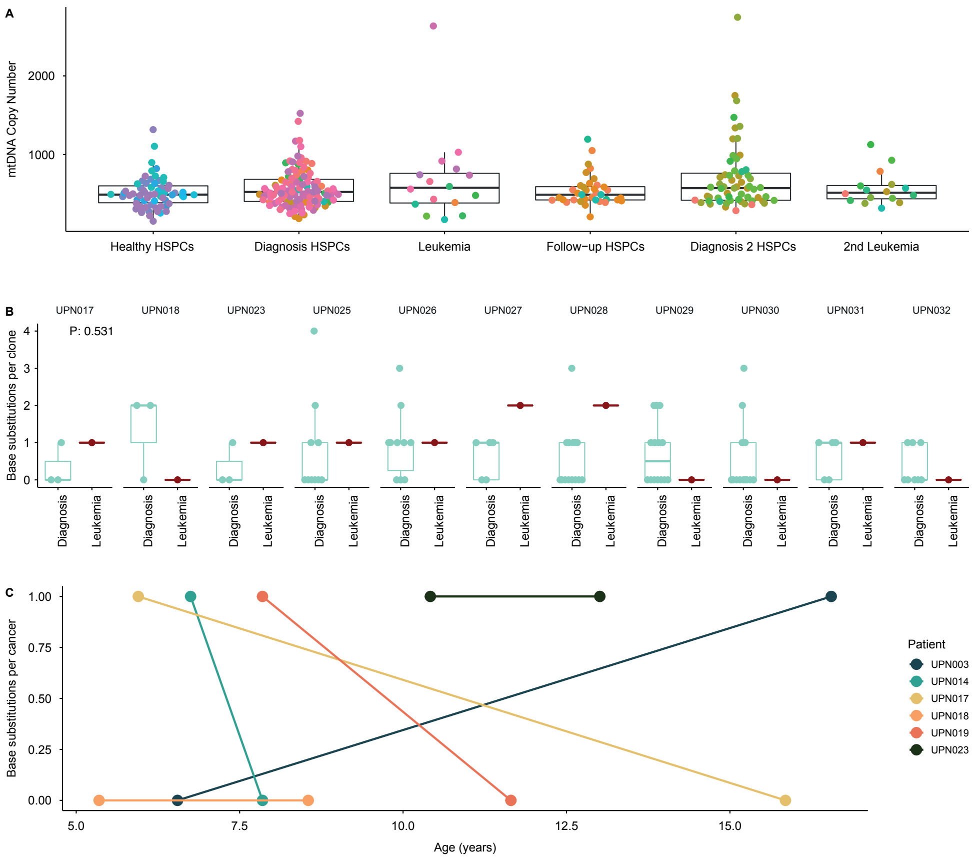


Figure S4. The effects of disease and treatment on mtDNA copy numbers. Related to Figure 4.
a Comparison of the mtDNA copy numbers between HSPCs from healthy donors (62 samples; 14 donors), HSPCs at diagnosis (107 samples; 17 donors), pediatric leukemias (16 samples; 16 donors), HSPCs from follow-up samples of patients in remission (36 samples; 8 donors), HSPCs at a secondary diagnosis (59 samples; 11 donors) and secondary pediatric leukemias (16 samples; 16 donors). The color of the dots indicates the donor. A small amount of jitter was added to the dots to prevent them from completely overlapping. **b** Comparison of the number of base substitutions between the primary leukemias and the normal HSPCs at diagnosis from the same patients. The patients are indicated at the top. The following patients were included: UPN017 (2 mutations; 4 samples), UPN018 (4 mutations; 4 samples), UPN023 (2 mutations; 4 samples), UPN025 (9 mutations; 11 samples), UPN026 (11 mutations; 11 samples), UPN027 (5 mutations; 6 samples), UPN028 (10 mutations; 15 samples), UPN029 (10 mutations; 15 samples), UPN30 (7 mutations; 14 samples), UPN031 (4 mutations; 6 samples), UPN032 (3 mutations; 9 samples). For each patient one leukemia sample was included and the rest are normal HSPCs. **c** The number of base substitutions between primary leukemias and secondary leukemias from the same patients is plotted against the age of the patients in years. The color indicates the patient and lines match samples from the same patient. The following patients were included: UPN003 (1 mutations; 2 samples), UPN014 (1 mutations; 2 samples), UPN017 (1 mutations; 2 samples), UPN018 (0 mutations; 2 samples), UPN019 (1 mutations; 2 samples), UPN023 (2 mutations; 2 samples).

Chapter 5

Investigation of single-cell genomes at nucleotide resolution using the PTA Analysis Toolkit (PTATO)

Freek Manders[1,†], Sjors Middelkamp[1,†], Markus J. van Roosmalen[1], Eline Bertrums[1], Inge van der Werf[1], Evangelia Antoniou[2], Dirk Reinhardt[2], Jurrian de Kanter[1], Niels M. Groenen[1], Mark Verheul[1], Arianne M. Brandsma[1], and Ruben van Boxtel[1]

[1]Princess Máxima Center for Pediatric Oncology and Oncode Institute, Heidelberglaan 25, 3584CS Utrecht, The Netherlands
[2]Department of Pediatric Hematology and Oncology, University Hospital Essen, Essen, Germany
[†]These authors contributed equally

Abstract

Single-cell whole genome sequencing allows for the detection of rare somatic mutations that are missed by conventional bulk sequencing. Single-cell sequencing has historically had a low accuracy and sensitivity but combined with the recently developed primary template-directed amplification (PTA) technology some of these issues have been alleviated. Nonetheless, while this technique reduces amplification biases and allelic dropout rates in the sequencing data, it still introduces many artifacts due to the DNA amplification. Here, we present the PTA analysis toolkit (PTATO), which integrates multiple modes of evidence to accurately distinguish naturally acquired somatic mutations from artifacts. The toolkit can identify not only single base substitutions (SBSs), but also small insertions and deletions (indels), and structural variants (SVs). We validated PTATO on internal data from multiple donors as well as on external data. Finally, we illustrate how our toolkit can be used to gain novel biological insights by applying it on blood cells from an acute myeloid leukemia (AML) patient. We demonstrate that leukemic drivers can be timed in the lifetime of a patient, which occurred years before the diagnosis of the AML, and that AML blasts are still able to differentiate.

Introduction

Somatic mutations play an important role in several diseases, such as neurodegenerative disorders and cancer[1,2]. Detecting somatic mutations using single-cell whole genome sequencing (WGS) can provide several advances over bulk sequencing of a sample containing many cells[3–5]. With regular bulk sequencing only (sub)clonal mutations can be detected. Therefore, any mutation present in a particular cell, which has not substantially clonally expanded, will be missed by bulk sequencing. As a result, most of the somatic mutations in healthy bulk samples are missed[1,6]. Additionally, while the (sub)clonal mutations in a cancer sample can be identified with bulk sequencing, many of the mutations that occurred after the malignant transformation or last clonal sweep, which could have occurred much earlier, will be missed[7–11]. Although these mutations occur in only a limited number of cells, they can represent small subclones that may ultimately contribute to chemotherapy resistance or metastases[12,13]. Furthermore, because somatic mutations are inherited by a cells' progeny, they can be used to determine the phylogenetic relationships between cells. By performing single-cell WGS on different cells of the same individual, phylogenetic trees can be constructed[14–16]. On top of this, determining the mutation burden of normal cells allows for a direct comparison to the clonal mutations in a bulk cancer sample, which are the mutations present in the cancer cell-of-origin or the last cancer cell that performed a clonal sweep[1,17–21].

There are currently several techniques, which allow for studying genome-wide mutation profiles at the single-cell level[2,22]. One approach is to clonally expand adult stem and progenitor cells *in vitro* to obtain sufficient DNA for WGS analysis[14,15,17,20,23–27]. A conceptually similar approach is to sequence naturally occurring clonal patches in healthy tissues using low-input sequencing[1,21,28,29]. By subsequently selecting only the clonal variants, the mutations present in the parental cell that gave rise to the clonal expansion can be studied. A disadvantage of these methods is that they can only be used on cells with sufficient self-renewal capacity *in vitro* and/or *vivo*.

Another approach to obtain sufficient DNA of a single cell for WGS analysis, without the necessity for clonal expansion, is to artificially amplify the genome using multiple displacement amplification (MDA)[30,31]. However, this method suffers from amplification biases across different regions of the genome as well as allelic dropouts and therefore the accuracy in mutation detection is low[4,5]. A recent modification of the MDA protocol is primary template-directed amplification (PTA)[5]. This technique incorporates exonuclease-resistant terminators next to regular nucleotides, which results in amplification products that undergo limited subsequent quasi-linear amplification[5]. Consequently, a larger fraction of the initial products of the original DNA molecule is amplified, which results in a more even read coverage across the genome. The benefit of this approach is that it can be applied to differentiated and single cancer cells, which otherwise could not be assessed. PTA can help us understand tumor heterogeneity in cancer, but also other diseases as illustrated by a recent paper, which showed that the neurons of Alzheimer's patients have an increased mutation load compared to the neurons of neurotypical patients[32].

While PTA generates less artifacts than MDA and has a much more even genome coverage resulting in a higher sensitivity, there is still a substantial number (hundreds to thousands) of artifacts left across different types of mutations like SBSs, indels, and SVs[5,33]. Copy number profiles generated from PTA-based WGS data are also relatively noisy compared to regular bulk WGS of clonally expanded cells[5]. Because of the substantial number of PTA artifacts combined with noisy copy number profiles, regular post-calling variant filtration is insufficient to remove these artifacts, especially when investigating cells with a relatively small mutation burden, such as noncancerous cells.

Several approaches could potentially be used to distinguish true somatic mutations from PTA artifacts. First, the mutation type and sequence context surrounding artifacts is not random[33]. False positive SBSs, for example, show a clear mutation pattern of C>T substitutions outside of CpG islands. Furthermore, in a diploid locus, the variant allele frequency (VAF) of a somatic variant should match that of the surrounding constitutive variants. When both alleles are equally amplified this VAF should be 0.5 in a diploid region, but when this is not the case, then the VAF of a somatic variant should change similarly to that of the surrounding germline variants[33]. Finally, if a true somatic mutation is located closely to a germline variant, then read-backed phasing can be used[34]. This analysis takes into account that if the somatic and germline variant are located on the same allele, then any read overlapping both their positions should contain both variants. In contrast, if they are on different alleles, then no single read should contain both variants. Current tools are limited by only using some of the approaches described above, instead of integrating them to distinguish true variants from artifacts with maximum precision and sensitivity.

PTA-based WGS might also enable structural variant detection in single cells. The relative uniform sequencing coverage of PTA data and ability to identify germline SBSs give opportunities to analyze read depth, B-allele frequencies, split reads, and concordant read pairs. However, currently no specific SV filtering tools exist for PTA-based WGS data, while commonly used structural variants analysis workflows for general WGS data are not optimized for PTA, leading to suboptimal results.

Here, we created a comprehensive PTA Analysis Toolkit, which integrates multiple complementary lines of evidence, using a machine-learning model, to distinguish true somatic SBSs, indels, and SVs from artifacts with an unprecedented precision and sensitivity for single-cell data. By applying PTATO to both novel and previously published PTA-based WGS datasets, we showed that it can accurately distinguish true variants from artifacts. Furthermore, we applied PTATO on single hematopoietic cells of a pediatric AML patient to show that oncogenic drivers occurred years before the diagnosis of the disease and that the leukemic stem cells in the *NUP98–NSD1* AML were still able to differentiate.

Results

PTATO has several main workflows, which separately filter SBSs and indels, create quality control plots, and determine SVs and copy numbers (Fig. 1). To filter potential somatic SBS and indels, PTATO uses machine-learning to integrate read-backed phasing, with allelic imbalance, and a wide genomic context. First, PTATO performs read-backed phasing of somatic variants with nearby heterozygous germline variants and generates a score indicating whether the variant is likely to be true. Additionally, PTATO analyzes the allelic imbalance, by checking if the VAF of a potential variant is similar to the surrounding heterozygous germline variants. Next, PTATO uses a random forest model to calculate for each detected variant the probability that it is an artifact. This model uses several features, such as the 10 bases sequence context around a variant, the distance to the nearest gene body, and the allelic imbalance (Fig. 1) (Methods).

Finally, for each sample, the optimal cutoff for the random forest score is calculated. This calculation is done using variants that are highly likely to be artifacts or true variants based on read-backed phasing of somatic variants with nearby heterozygous germline variants. This approach ensures that the cutoff score is well calibrated for different samples with varying amplification qualities. Next to the calculated cutoff score, users can also use a custom cutoff score if a higher sensitivity or precision is required. PTATO makes it easy to choose a custom cutoff by calculating several statistics, such as the precision, and sensitivity for different possible custom cutoff scores.

Most PTA-based WGS runs lead to uniform genome coverage, but some samples may have a suboptimal genome amplification, leading to noisy copy number plots and many artificial variants. Therefore, PTATO first collects and visualizes a variety of sequencing quality control metrics that can aid in correct interpretation of PTA-based WGS data (Fig. S1).

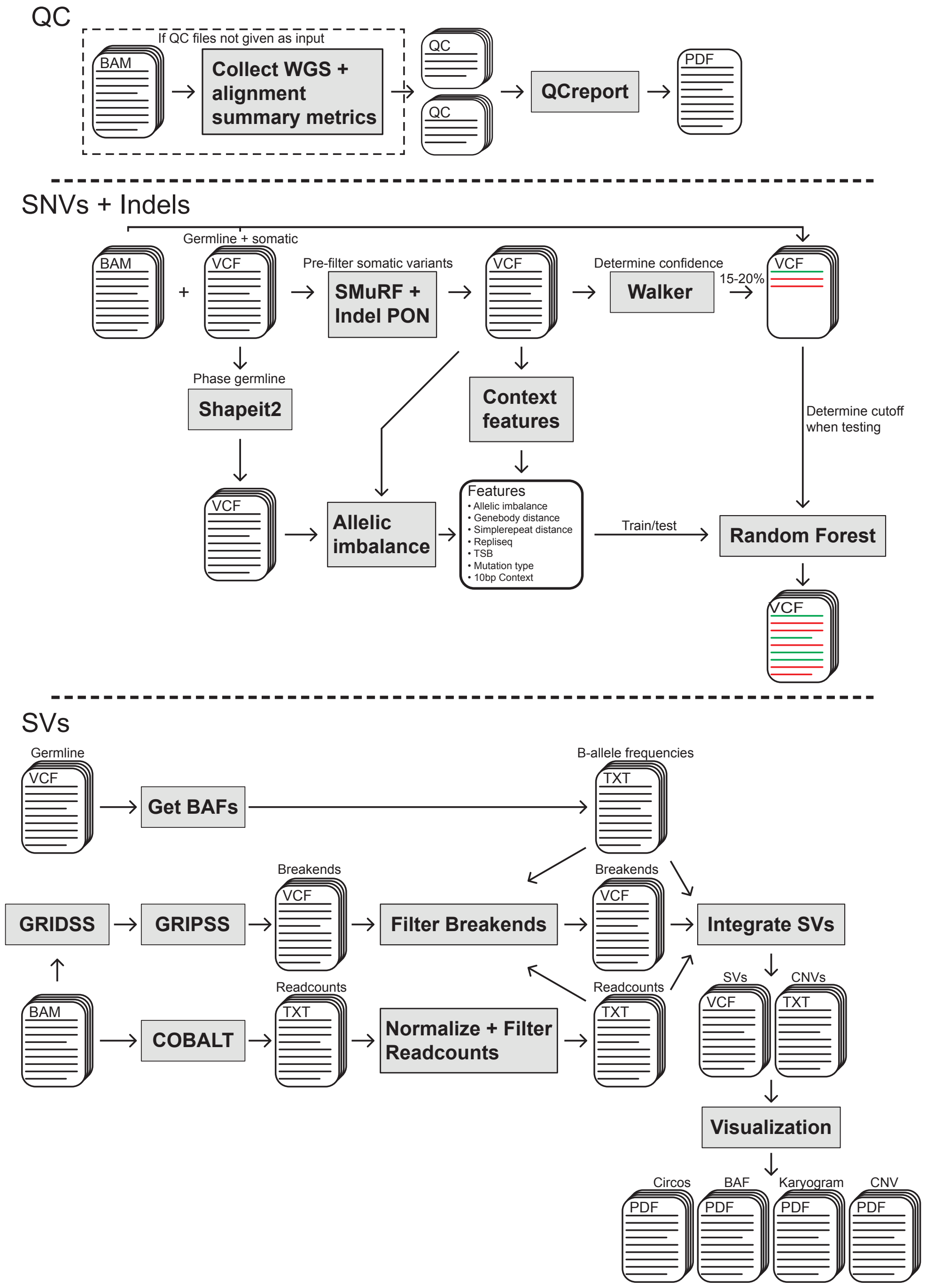


Figure 1: Schematic overview of the PTATO tool.
The steps taken by the PTATO tool are shown separately for the quality control steps, the SBSs/indels, and the SVs.

PTA-based WGS samples have a relatively uniform genome coverage

To optimally train the SBS and indel random forests, we needed a balanced set of true positive and false positive variants for both variant types. Any somatic variant that is shared between a bulk cancer sample and a single cell of the same cancer is likely to be true. Therefore, we sequenced single leukemic cells using PTA as well as a bulk AML sample and a germline control sample (mesenchymal stromal cells) from three AML patients, two of which had Fanconi anemia and one with a *NUP98-NSD1* translocation (Table. S1). Variants shared between the PTA-based samples and the bulk AML sample, but absent in the germline control, were considered high-confidence somatic variants and used as true positives in the training set. To increase the variety of our training data we created a *FANCC* knockout cell line using CRISPR/Cas9 gene editing and sequenced both a clone and a subclone, using a previously described method[35]. The subclone was sequenced twice. Once using regular bulk sequencing and once using PTA. Any variants that were present in both subclone samples, but not in the clone are somatic mutations that were acquired after the *FANCC* gene was knocked-out and were used as true positives in the training set.

Unique variants that were likely artifacts based on read-backed phasing were used as false positives (Methods). To increase our set of false positive variants, we clonally expanded a cord-blood hematopoietic stem cell (HSC), and sequenced three daughter cells. We detected 41 shared and 1278 unique variants. Since cord blood cells only contain around 40 true mutations[18], most of the unique mutations are artifacts. Therefore, all unique variants in these sample were used as false positives.

PTA-based WGS samples had an average mean read depth of 17.3 (+-1.62 s.d.), with on average 90.2% (+-2.24 s.d.) of the genome having a read coverage of at least 5x, showing that the reads were well distributed across the genome.

PTATO performs well on SBSs and indels

To validate that a low read-backed phasing score is indeed indicative for a PTA-artifact, we compared the phasing scores of variants shared between the PTA-based and the bulk WGS samples (true variants) and the mutations unique for each PTA-based sample (mix of true and false positives). The shared variants had a higher median phasing score compared to unique variants for SBSs ($P < 2.2*10^{-16}$, Wilcoxon test), indicating that the phasing score can accurately distinguish true variants from artifacts (Fig. 2a). For indels the difference was similar but not significant ($p = 0.252$, Wilcoxon test), since only 11 shared indels were phased (Fig. S2a). As expected, the sets of unique calls also contain variants with high walker scores, because, in addition to the PTA artifacts, these sets also contain true mutations. The presence of true mutations in the set of unique variants is exemplified by the relatively high walker scores of unique SBSs of the *FANCC* knockout, which was expected to have a high fraction of true mutations, because of defective DNA repair.

To train the random forest model, we used 756 artifact and 756 true positive SBSs. For the indel random forest we used 758 artifact and 55 true positive indels. The out-of-bag error rate of the random forest was 26.39% for the SBSs, which showed that the random forest model can filter out artifacts relatively well (Fig. S2b). We also determined how the precision and recall of the random forest model changes with the cutoff and observed an area-under-the-curve of 0.79 (Fig. 2b), showing that the model is relatively accurate. The indel random forest model performed better with an out-of-bag error rate of 3.57% and an area-under-the-curve of 0.848 (Fig. S2c, d). Because of the large class imbalance, the standard non-optimized cutoff results in almost 50% of true positive variants being filtered, however a higher specificity can be obtained with a less stringent cutoff, as can be seen in the receiver-operator curve (Fig. 2b). To further validate the performance of the random forest models, we examined the mutational patterns of the artifact and true positive datasets. The artifact SBS spectrum shows a very strong enrichment for C>T mutations at non-CpG sites, as previously described (Fig. 2c)[33], while the artifact indel spectrum predominantly showed single C and T insertions within C and T mononucleotide repeats (Fig. S2e). The 96-trinucleotide patterns of the variants predicted to be false or true by the random forest model were very similar to the artifact and true positive input variants with cosine similarities of respectively, 0.986 and 0.966 for SBSs and respectively 1 and 0.772 for indels, showing that the random forest model accurately recapitulates the mutational patterns of their input samples (Fig. 2d; Fig. S2f). Next, we fitted the PTA-artifact and HSPC mutational signatures to the input data before and after filtering with the SBS random forest. The HSPC signature is a signature present in hematopoietic stem and progenitor cells[18], while the PTA signature is a previously published pattern of PTA artifacts[33]. The contribution of mutations with the PTA artifact signature was strongly reduced after filtering, while the contribution of the HSPC signature showed only a small decrease (Fig. 2e). Since no PTA signature has been reported for indels so far, we used the COSMIC signatures for this case. The input data contained a large contribution of ID1, ID7 and, ID16 which was mostly removed by the random forest (Fig. S2g). It is difficult to distinguish true C and T insertions within repeat regions from artifacts since the latter are so much more abundant. Therefore, we filtered these variant calls out for subsequent analyses.

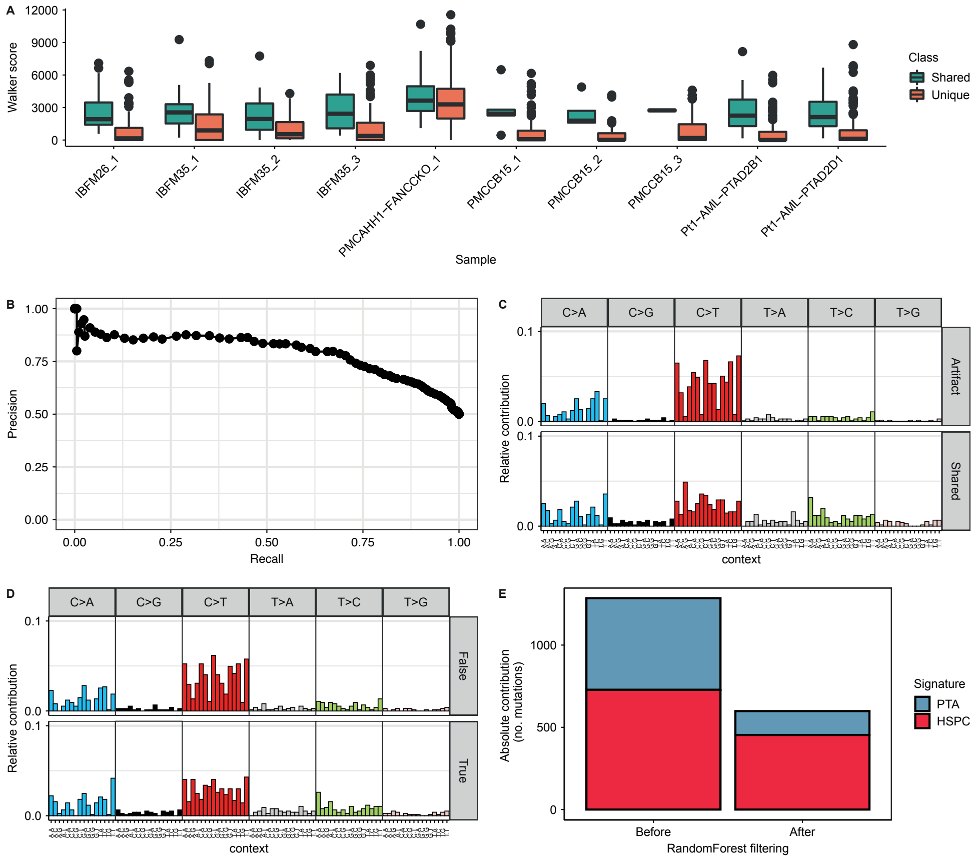


Figure 2: The walker and random forest perform well on the training data.
a Boxplot showing the walker scores of the SBSs in the samples used to train the random forest. The color indicates if mutations are shared between samples or are unique. Unique mutations with a walker score below 1 were used to train the random forest. **b** Precision and recall curve showing the performance of the random forest using all input variables on the out-of-bag training data for cutoffs between 0 and 1 with a step of 0.01. **c** Relative contribution of each trinucleotide change to the point mutation spectrum for the mutations used to train the random forests separated into the artifact variants (n = 756; samples = 10) and the mutations that were shared between samples (n = 756; samples = 10). **d** Relative contribution of each trinucleotide change to the point mutation spectrum for the mutations predicted to be artifacts (n = 745; samples = 10) and the mutations predicted to be true somatic mutations (n = 767; samples = 10). **e** Absolute contribution of each mutational signature for all the mutations used to train the random forests (n = 1512; samples = 10) and for all the mutations predicted to be true (n = 767; samples = 10) by the random forests.

PTATO works robustly on external data

After validating PTATO on our own in-house data, we tested its robustness by applying it on external data from the original PTA publication[5], as differences in sample handling and sequencing logistics could potentially impact the performance of our method. We performed variant calling and then applied PTATO on the sequencing data of human umbilical cord blood PTA samples treated with either N-ethyl-N-nitrosourea (ENU), D-mannitol (MAN), or a vehicle (VHC) control[5]. Samples treated with either a moderate or high concentration of ENU clearly had an increased mutation load, compared to the other samples, in accordance with the original study (Fig. 3a). Since this increased mutation load is likely caused by true mutations, the ratio of true mutations is likely also higher in these samples. Indeed, PTATO filtered out a lower percentage of variants in the samples treated with a high concentration of ENU with the median mutation load being reduced by 23.9% from 5574 to 4240, whereas the median mutation load in VHC treated samples was reduced by 64.4% from 1159 to 413. In comparison, the mutation loads found by the original PTA paper were quite a bit higher, with the high concentration ENU samples having a median mutation load of around 4700 and the VHC samples having a median mutation load of around 800[5]. This shows that PTATO filters more stringently than the original PTA paper. The sample with the lowest percentage of the genome covered by at least one read also had the lowest mutation load among the samples treated with a moderate concentration of ENU (Fig. 3a; Fig. S1). The measured mutation load is thus likely an underestimation, showing that the quality control plots generated by PTATO can be used to identify potentially bad quality samples. Next, we looked at the mutation patterns of the cord blood samples before and after filtering. The C>T peaks associated with PTA artifacts were less pronounced after filtering and the contribution of the PTA signature, but not other signatures, was strongly reduced (Fig. 3b, c; Fig. S3a, b). Additionally, the samples treated with ENU had a large contribution of a previously experimentally defined ENU-induced SBS signature[36], indicating that PTATO can accurately recover the patterns of mutations present in a sample.

Neither the MAN or ENU treatments caused an increased burden of indels, which is in line with earlier studies that showed that ENU does not cause indels[36] (Fig. S4a). The mutational patterns consisted of mostly C and T insertions before filtering and T deletions after filtering (Fig. S4b). These deletions match signature ID2, which is associated with polymerase slippage during DNA replication (Fig. S4c). In contrast to the data used to train PTATO, the unfiltered data contained more T than C insertions. The lower abundance of T insertions in our data could be caused by improvements made to the PTA protocol, however, this is not certain. Overall, these data indicate that it might be beneficial for users of PTATO to re-train the PTATO random forests with their own data.

Accurate detection of structural variants in PTA-based sequencing data

Structural variants can also be detected in PTA-based WGS data, but optimized workflows to detect and filter such variants in this type of data are currently lacking.

Existing bioinformatic tools for single-cell genome sequencing are usually developed for low-coverage sequencing and are limited to detection of copy number changes based on read depth and do not use other modes of information, such as split reads and concordant read pairs[37,38]. More comprehensive SV calling pipelines exist for regular bulk WGS data, but they are hampered by many false positive variants in PTA-based WGS data (Fig. 4a; Fig. S5).

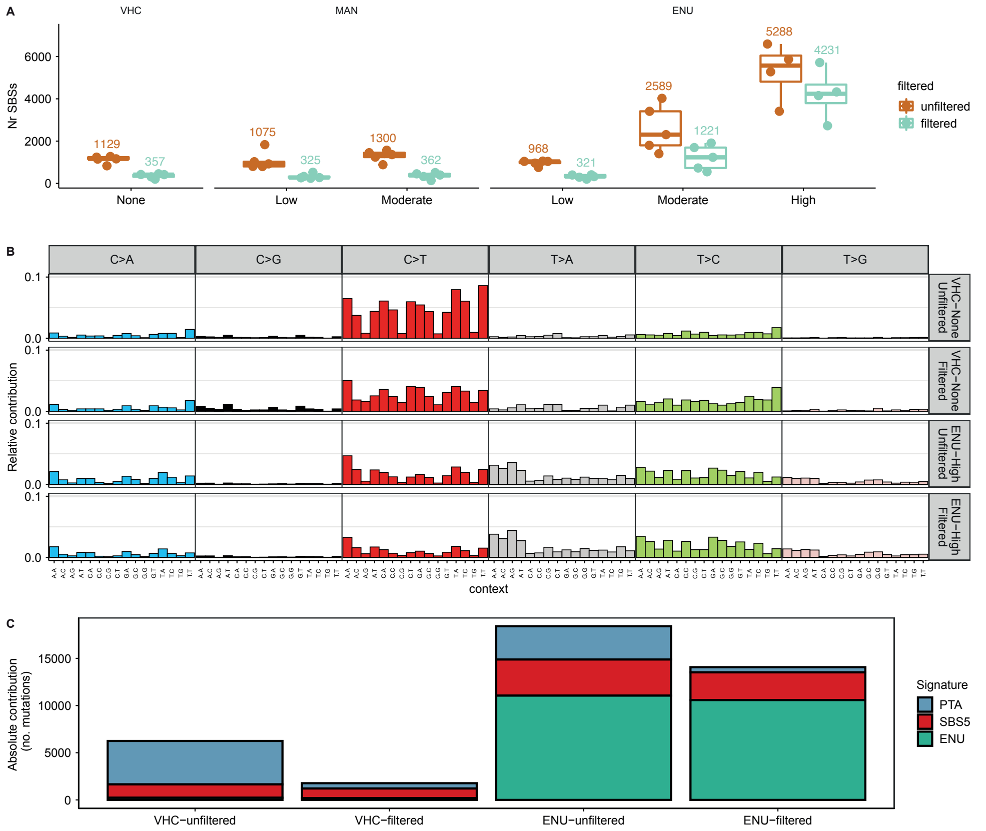


Figure 3: PTATO accurately filters SBSs in external cord blood samples.
a Boxplot showing the number of SBSs for human cord blood samples treated with different concentrations of a vehicle control (VHC; n = 5), D-mannitol (MAN; low: n = 5, moderate: n = 5) or N-ethyl-N-nitrosourea (ENU; low: n = 5, moderate: n = 5, high: n = 4). The color indicates if the mutations were filtered by PTATO. **b** Relative contribution of each trinucleotide change to the point mutation spectra of the cord blood samples treated with a vehicle control (unfiltered: n = 5646; filtered: n = 1787; samples = 5) or with a high concentration of ENU (unfiltered: n = 21152; filtered: n = 16923; samples = 4) before and after filtering with PTATO. The mutations of samples with the same treatment were pooled together. **c** Absolute contribution of mutational signatures PTA, ENU, and SBS5 to the cord blood samples treated with a vehicle control (unfiltered: n = 5646; filtered: n = 1787; samples = 5) or with a high concentration of ENU (unfiltered: n = 21152; filtered: n = 16923; samples = 4) before and after filtering with PTATO. The mutations of samples with the same treatment were pooled together.

To enable accurate detection of all types of structural variation in PTA-based data, PTATO applies several SV calling, filtering, and normalization steps, using a combination of existing tools and custom scripts that make optimal use of the data's nucleotide resolution (Fig. 1) (Methods). While the read depths of PTA-based WGS data are noisy compared to bulk WGS data, many of the fluctuations in coverage are highly recurrent between PTA-based WGS samples (Fig. S6a). For example, three single HSPCs from a donor with Fanconi anemia had a mean cosine similarity of 0.94 when comparing their read counts in 1kb bins across the genome (Fig. S6b). In contrast, they had a mean cosine similarity of only 0.77 when compared to an AML bulk sample from the same donor. This recurrence is used to smoothen the read counts of each sample, after which they are binned (Fig. 4b; Fig. S6c). Subsequently, each copy number call is supported using the VAFs of heterozygous germline SBSs. Additionally to identify SVs with a nucleotide resolution, PTATO detects and filters breakpoints and single breakends, which are breakpoints for which only a single end was detected. We found that many false positive SVs occurred in multiple unrelated individuals, enabling additional filtering based on recurrency. Furthermore, we noted that many called SV candidates in PTA-based data appear to be small duplications and inversions (<1kb) with only one breakpoint junction, which may correspond to small chimeric DNA molecules generated during the PTA reaction (Fig. S5b)[39]. By filtering these events PTATO removes most false positive calls (Fig. S5a). Finally, PTATO integrates the copy number segments, BAF segments, and filtered breakends to obtain a high confidence set of structural variants.

To optimize our SV filtering strategy, we used the three single HSPCs of one of the donors with Fanconi anemia. Comparison of the SVs in the single HSPCS with SVs detected in the bulk-sequenced AML sample shows that the blood of this patient is highly clonal. After filtering the PTA-based data of the HSPCs with PTATO, most structural variants that were present in the AML bulk sample, could still be detected while the number of false positive calls was greatly reduced (Fig. 4b, c, d). The detected SVs included a t8;21 translocation, which causes the *RUNX1-RUNX1T1* fusion, which is a known cancer driver[40]. Several SVs present in the blasts are not detected in the HSPCs, indicating that these HSPCs are pre-leukemic cells (Fig. 4b, d). For example, the AML blasts show a gain of chromosome 13, which is not present in any of the HSPCs, suggesting that this chromosomal gain could be an additional driver in this AML (Fig. 4b, c). HSPC3 also shows a partial loss of chromosome 2, which is not detected in any of the other samples.

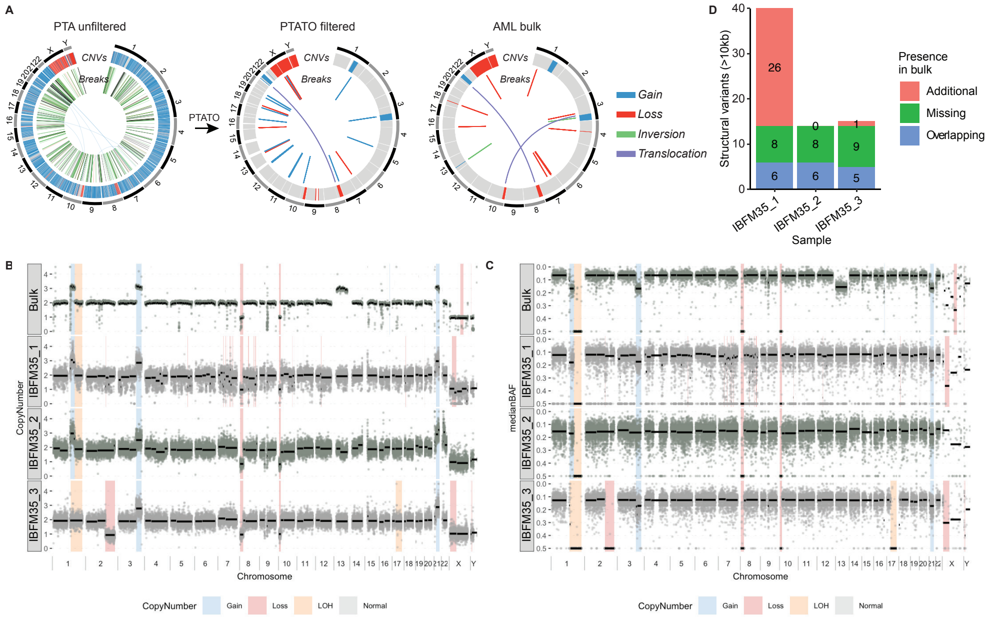


Figure 4: PTATO enables accurate structural variant detection in PTA data.
a Circos plots showing structural variants before (left) and after (center) PTATO filtering in the PTA-based IBFM35_2 sample and in the corresponding bulk AML (right). The color indicates the type of structural variant. **b** Copy number profiles of an AML bulk sample and three PTA-based samples of HSPCs from a donor with Fanconi Anemia. Each dot represents the mean copy number in a 100kb bin. The background highlights show the final copy number state called by PTATO. Sample IBFM35_1 shows many small deletions on chromosomes 7 and 8, likely due to uneven PTA amplification. The amplification of chr13 is only present in the AML cells, suggesting that the HSPCs are still pre-leukemic cells. **c** BAF-like profiles showing the mean deviation of allele frequencies (DAFs) of germline SNVs in 100kb bins of the same samples as in b. PTATO only calls copy number changes if they have coverage and DAF support, as shown by the highlights in this figure. The last part of chromosome 1 contains a loss-of-heterozygosity (DAF of 0.5) without a change in copy number. **d** Quantification of SVs detected by PTATO that are shared or not between the AML bulk sample and the PTA-based HSPCs.

Abundance of C>A mutations caused by a continuous mutational process

One major benefit of PTA followed by PTATO analysis is that it can be applied to study the genomes of single cells that cannot be easily clonally expanded either *in vitro* or *in vivo*. We previously observed a subset of pediatric AML cases that had an above average mutation load and an abundance of C>A mutations, which have been attributed to oxidative stress-induced mutagenesis[19,36]. These latter AML patients also had a better overall survival compared to patients with a lower mutation load. We hypothesized that the abundance of C>A mutations might have been generated by a single mutational burst caused by the myeloid differentiation[41]. To test this hypothesis, we sequenced 6 differentiated cell types as well as a HSC and multipotent progenitor (MPP) cell of one of these pediatric AML patients, which had a *NUP98-NSD1* fusion as well as *IDH2*, *WT1*, and *FLT3* driver mutations[19]. PTATO removed a large fraction of mutations in all samples except for the AML bulk, which is not a PTA sample and thus does not contain any PTA artifacts (Fig. S7). PTATO retained 93.8% of all SBSs in this sample, which shows its high sensitivity. The differentiated cells as well as the MPP, HSC, and AML cells all contained similar autosomal mutation patterns with respect to both the SBSs and indels, showing that the abundance of C>A mutations was not specific to the AML (Fig. 5a; Fig. S8; Fig. S9a, b).

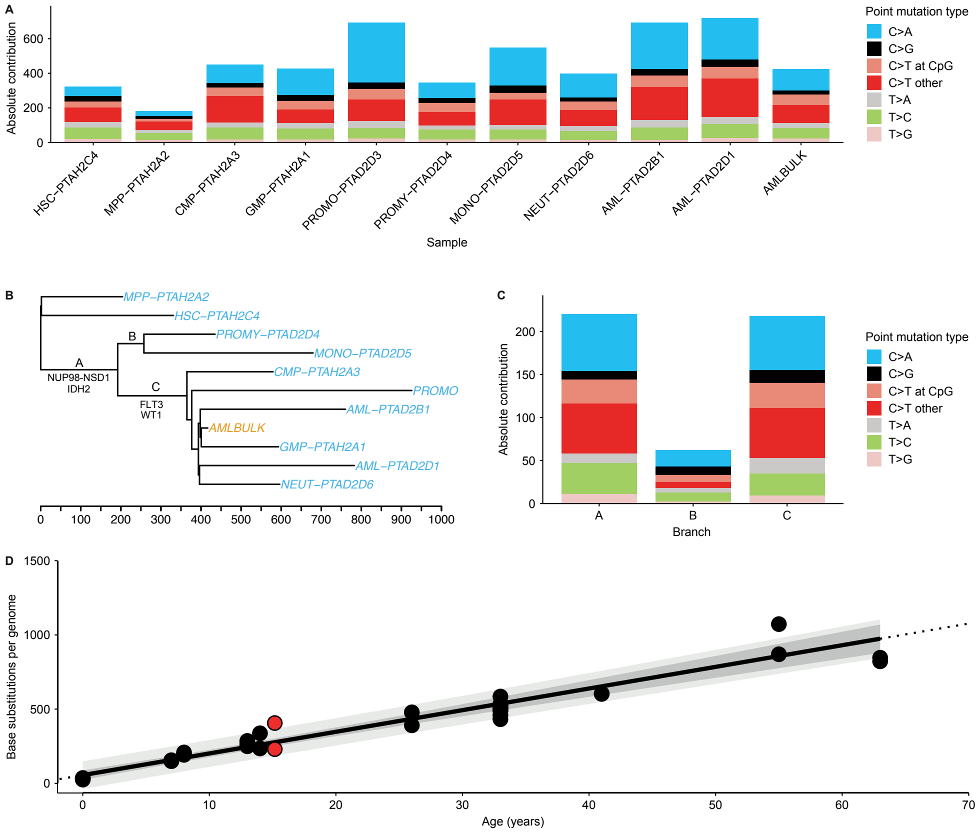


Figure 5: PTA samples can be used to make a lineage tree, allowing for the timing of drivers.
a Spectrum of the six types of base substitutions for different cell types in an AML patient. The colors indicate the type of base substitution. **b** Phylogenetic lineage tree of an AML patient. Each tip represents a single cell that was sequenced with PTA or the AML bulk. The length of the branches indicates the number of somatic bases substitutions in that branch of the tree. Blue indicates that a sample is a PTA sample and orange that a sample is an AML bulk. **c** Spectrum of the six types of base substitutions for the three branches indicated in b. The colors indicate the type of base substitution. **d** The number of base substitutions is plotted against the age in years of the donor. Each dot is a sequenced sample. The black line shows the mean fitted number of substitutions at that age (linear mixed-effects model). The dark grey area shows the 95% confidence interval of the model, while the light grey area shows the 95% prediction interval. Black dots are clonally expanded HSCPs from normal donors (n = 33; donors = 11), whereas the two red dots are an HSC (top) and MPP (bottom) from the AML patient.

Additionally, we found that several cells shared many mutations with each other and the AML bulk (Fig. S10). Based on the shared SBSs we generated a phylogenetic lineage tree showing the lineage relationships between the cells and the location of the drivers that had been previously identified in this AML (Fig. 5b). We found that the *NUP98-NSD1* fusion and *IDH2* driver mutation occurred early in the tree. A subset of the cells with these drivers also shared the *WT1* and *FLT3* driver mutations with the AML, indicating that these drivers occurred later during leukemogenesis. This observation is in line with previous studies, which found that *IDH2* mutations are generally acquired early while *FLT3* mutations are often acquired later during leukemogenesis[42–45]. The *WT1* and *FLT3* drivers occurred in the tree before the AML bulk WGS-sample. This indicates that they were necessary for the development of the AML and that the *NUP98-NSD1* translocation and IDH2 driver were not sufficient. The C>A mutations were already abundant in the earliest branches of the phylogenetic tree, irrespective of the presence of any AML drivers (Fig. 5c). This shows that the C>A mutations did not occur in a single burst but were instead generated by a continuous mutational process.

AML drivers occur years before diagnosis

Based on the mutation accumulation of HSPCs during healthy life[18,46], we estimated that the *NUP98-NSD1* fusion and *IDH2* driver occurred before the patient was 9.12 years old[18,46] (Methods). Since drivers might increase the mutation rate, this age could be an overestimation. Therefore, as an orthogonal approach, we divided the number of mutations in the branch containing the drivers by the average of the total number of mutations in the cells that contained the first set of drivers but not the *WT1* and *FLT3* drivers. This method suggested that the first set of drivers occurred before the patient was 7.47 years. However, this might be an underestimation, because cancer drivers could have increased the mutation rate after they occurred, causing later branches to be elongated. Both these approaches estimate that the drivers occurred at least years before the diagnosis of the AML at 15.17 years, matching adult cancers which were also found to contain drivers years before diagnosis[7–11]. When applying these same methods on the *WT1* and *FLT3* drivers we estimated that they occurred before the patient was 19.9 or 11.81 years old, confirming that the first method overestimates the age, while the second method likely underestimates the age. These observations suggest that the latter drivers likely occurred relatively close to the diagnosis of the AML, further indicating that they were necessary for the development of the AML.

The clade containing the *WT1* and *FLT3* drivers included both differentiated and progenitor cells, such as neutrophils and promonocytes, which were genetically closely related to the AML samples. Since these differentiated cells contain the mutations driving the AML, they could be differentiated blasts, which would match a previous finding that AML blasts can differentiate[47]. Interestingly, while the AML blasts were closely related to the differentiated cells, based on the phylogenetic tree, they seemed to have a higher mutation load. However, a larger sample size from multiple donors is needed to confirm this.

To further validate PTATO, we compared the autosomal mutation load of the hematopoietic stem cell (HSC) and multipotent progenitor (MPP) cell, neither of which contained the AML drivers, with the mutation load of HSPCs that were clonally expanded and sequenced without PTA[18,46]. After correcting for differences in the surveyed area of the genome, the MPP had a SBS mutation load that matched the expected mutation load for a donor of this age, while the HSC had a slightly higher mutation load (Fig. 5d). When comparing the mutation load of the indels, both the HSC and MPP had, after filtering with PTATO, a mutation load that was within the expected range (Fig. S9c). This shows that PTATO can identify the somatic mutation loads of samples relatively accurately, although it may lead to slight overestimations.

Discussion

Here we have shown that PTATO can accurately identify SBSs, indels, and SVs in PTA-based WGS data, while effectively filtering out artifacts. During the last few years, the mutations of the normal stem and progenitor cells of many tissues have been characterized, allowing the determination of mutation rates during healthy aging and identifying the underlying mutational processes[1,2,16,18,20,21,24,25,29,48,49]. With the use of single-cell DNA sequencing, these analyses could be expanded beyond the stem cell compartment to include many different cell types, leading to a better understanding of the mutagenesis of normal cells and a better understanding of tumor heterogeneity. PTA-based WGS can be used to investigate the somatic evolution of a cancer after its initiation by the original transformed cell or after the last clonal sweep, as shown by our finding that AML blasts were able to differentiate in the patient we investigated.

Users can use the random forests generated here, but can also generate their own random forests, which might give better results when their data is generated differently. For example, if the errors in PTA data change with future updates to the PTA protocol, then the random forests in PTATO can also be easily updated. Next to the PTA probability cutoff calculated by PTATO, users can also use their own cutoffs, which makes the toolkit useful for analyses that have specific sensitivity and specificity requirements.

Next to PTA, other methods for detecting non-clonal somatic variants in non-dividing cells, such as single-molecule duplex sequencing, META-CS, and MALBAC have been developed[50–52]. While this first method is highly accurate, it detects mutations at the level of single-molecules instead of single-cells and can therefore not identify which mutation is present in which cell, making it impossible to perform phylogenetic analyses[50]. META-CS, which separately amplifies both strands of the DNA, is also highly accurate, however its sensitivity is limited, because it only covers around 50% of the genome[51]. MALBAC, which performs quasi-linear preamplification using looped amplicons, has a higher sensitivity, but it also generates around 100,000 false positives[52]. Overall, PTA is currently the only single-cell whole genome amplification method that combines a high sensitivity and precision.

Here we have tested PTATO on PTA data, however the different types of information that PTATO uses to distinguish artifacts from true variants are not specific for PTA-based WGS data. PTATO could also be applied on other *in vitro* whole genome amplification methods, like MDA and MALBAC.

Next to PTATO, LiRA and SCAN2 can also be used to analyze PTA data[33,34]. LiRA uses read-backed phasing to identify SBSs in PTA data. While this tool has a high precision, it can only detect the +-27% of SBSs that are located near to heterozygous germline variants[34]. The tool also cannot detect indels or SVs. SCAN2 can detect both SBSs and indels by using the allelic imbalance and trinucleotide context of potential mutations to distinguish true variants from artifacts. However, by only focusing on the trinucleotide context of potential mutations it is ignoring the impact that the wider genomic context has on mutations[53–55]. Furthermore, the tool does not make use of read-backed phasing, cannot detect SVs, and has a sensitivity of only 45.7% for SNVs[33]. In contrast to these tools, PTATO uses machine-learning to combine read-backed phasing, allelic imbalance, and a wider genomic context, in order to detect SBSs, indels, and SVs with maximum precision and sensitivity.

Limitations

PTATO performs best at removing SBS and indel artifacts in diploid regions of the genome, because copy number gains and losses can influence the walker score and the allelic imbalance. While PTA combined with PTATO has an unprecedented accuracy and sensitivity at directly sequencing single cells, it is still not as accurate as sequencing cells that were clonally expanded. However, for differentiated- and other cells that cannot be sequenced via clonal expansions, PTA with PTATO provides an important advance over existing methods. In general, PTA data has a good quality and shows an even coverage over the genome. However, it is still important to perform quality control checks to remove the occasional low-quality sample, because some samples can have artifacts caused by DNA damage or an improperly amplified allele of one or more chromosomes. The quality control figures generated by PTATO can help determine the quality of a sample.

With PTA it is possible to directly analyze whole genomes of single cells at an unprecedented accuracy and sensitivity. PTATO allows for this data to be easily analyzed, and for the removal of artifacts, further unlocking the potential of PTA. We expect that PTATO will be used to further our understanding of genomes at the single-cell level.

Material and methods

Human bone marrow biopsies and umbilical cord blood

The bone marrow sample of patient Pt1 was obtained via the biobank of the Princess Máxima Center for Pediatric Oncology with ethical approval under proposal PMC-LAB2018-007. Written informed consent for this included individual was obtained by the Princess Máxima Center. The use of material for this study was approved by the Biobank and Data Access Committee of the Princess Máxima Center. Additionally, the umbilical cord blood sample of donor CB15 was obtained via the University Medical Center Utrecht (UMCU). The collection of cord blood samples was approved by the Biobank Committee of the UMCU (protocol number 19-737). Informed consent for these samples was obtained by the UMCU. Furthermore, the samples from IBFM26 and IBFM35 were obtained from the German Society of Pediatric Oncology and Hematology (GPOH) via a material and data transfer agreement. Informed consent for these samples was obtained by the GPOH.

Cell Isolation and Flow Cytometry

Bone marrow mononuclear cells and cord blood derived cells were stained for FACS after thawing. The following combinations of cell surface markers were used to define cell populations:

HSCs: Lin−CD11c−CD16−CD34+, CD38−,CD45RA−, CD90+; MPPs: Lin−CD11c−CD16−CD34+, CD38−,CD45RA−, CD90-; CMPs: Lin-CD11c-CD16-, CD34+, CD38+ and CD45RA-; GMPs: Lin-CD11c-CD16-, CD34+, CD38+ and CD45RA+; Promyelocytes: CD117+, HLADR-; Neutrophils: CD117-HLADR-CD11b+CD14-; Promonocytes: CD117-HLADR+CD15+CD14-; Monocytes: CD117-HLADR+CD15dimCD14+. AML blasts were selected based on diagnostic immunophenotyping data if available. In most cases, these blasts were CD33, CD38, and/or CD34 positive. Cells were single cell sorted on an SH800S Cell Sorter (Sony).

FACS and western blot antibodies

Antibodies used for cell sorting were as follows: CD117-BV421 (clone YB5.B8, 1:50) was obtained from BD. All FACS antibodies were obtained from BioLegend. CD34-BV421 (clone 561, 1:20), lineage (CD3/CD14/CD19/CD20/CD56)-FITC (clones UCHT1, HCD14, HIB19, 2H7, HCD56, 1:20), CD38-PE (clone HIT2, 1:50), CD90-APC (clone 5E10, 1:200), CD45RA-PerCP/Cy5.5 (clone HI100, 1:20), CD33-PE/Cy7 (clone WM53, 1:20), CD16-FITC (clone 3G8, 1:100), CD11c-FITC (clone 3.9, 1:20), HLADR-FITC (clone L243, 1:20), CD14-AF700 (clone HCD14, 1:50), CD11b-APC (clone IRFC44, 1: 20) and CD15-PE (clone HI98, 1:20). WESTERN Abs.

Generation of gene knockouts in AHH-1 cell lines

Human B-lymphocyte AHH-1 (CRL-8146) cells were purchased from ATCC. Cells were cultured in RPMI (Roswell Park Memorial Institute) 1640, GlutaMAX medium (Gibco, Thermofisher, US) supplemented with 1% Penicillin-Streptomycin (PenStrep, Gibco, Thermofisher, US) and 10% horse serum (HS, Gibco, Thermofisher, US). Guide RNAs (FANCC: 5’-GCAAGAGATGGAGAAGTGTA-3’ and MSH2: 5’-GTGCCTTTCAACAACCGGTTG-3’) were cloned into pSpCas9(BB)-2A-GFP (PX458) vector (Addgene#48138). AHH-1 cells were transfected using Lipofectamine 2000 (Thermo Fisher Scientific). One to two days after transfection, GFP-positive transfected cells were single-cell sorted for clonal expansion on a SH800S Cell Sorter (Sony), which was also used for subsequent clonal steps. MSH2 or FANCC gene knockout status was confirmed using western blot, Sanger sequencing and whole genome sequencing. For the MSH2 knockout clonal line, a second clonal step was performed at day 48 after the first clonal step and a third clonal step at day 36 after the second clonal step. PTA was performed 47 days after the third clonal step. For the FANCC knockout clonal line, a second clonal step was performed 58 days after the first clonal step, and PTA was performed 56 days after the second clonal step. Cells were harvested for DNA extraction when the cell lines were sufficiently expanded after the clonal steps.

PTA whole genome amplification and whole genome sequencing

PTA whole genome amplification of single cells was performed according to the manufacturer’s protocol (BioSkryb Genomics). Instead of 10 minutes cell lysis on ice as indicated in the protocol, lysis was performed by 5 minutes incubation on ice followed by 5 minutes incubation at room temperature to maximize DNA denaturation as previously described[56]. DNA from bulk AML and germline control samples (MSCs or T-cells) was isolated using the DNeasy DNA Micro Kit (QIAGEN) or DNeasy Blood & Tissue Kit (QIAGEN) according to the manufacturer’s instructions. WGS libraries were generated using standard protocols (Illumina). Libraries were sequenced to 15-30x genome coverage (2x150bp) on an Illumina NovaSeq 6000 system at the Hartwig Medical Foundation (Amsterdam, the Netherlands).

Primary processing WGS data
WGS reads were mapped against the human reference genome (GRCh38) using the Burrows-Wheeler Aligner (v0.7.17) mapping tool with settings ‘bwa mem –c 100 –M’[57]. Sequence reads were marked for duplicates using Sambamba v0.6.8. Realignment was performed using the Genome Analysis Toolkit (GATK) (v4.1.3.0)[58]. A description of the complete data analysis pipeline is available at: https://github.com/ToolsVanBox/NF-IAP (v1.3.0).

SBS and indel variant calling
Raw variants were multisample-called by using the GATK HaplotypeCaller and GATK-Queue with default settings and additional option ‘EMIT_ALL_CONFIDENT_SITES’. The quality of variant and reference positions was evaluated by using GATK VariantFiltration with options: “--filter-expression 'QD < 2.0' --filter-expression 'MQ < 40.0' --filter-expression 'FS > 60.0' --filter-expression 'HaplotypeScore > 13.0' --filter-expression 'MQRankSum < -12.5' --filter-expression 'ReadPosRankSum < -8.0' --filter-expression 'MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)' --filter-expression 'DP < 5' --filter-expression 'QUAL < 30' --filter-expression 'QUAL >= 30.0 && QUAL < 50.0' --filter-expression 'SOR > 4.0' --filter-name 'SNP_LowQualityDepth' --filter-name 'SNP_MappingQuality' --filter-name 'SNP_StrandBias' --filter-name 'SNP_HaplotypeScoreHigh' --filter-name 'SNP_MQRankSumLow' --filter-name 'SNP_ReadPosRankSumLow' --filter-name 'SNP_HardToValidate' --filter-name 'SNP_LowCoverage' --filter-name 'SNP_VeryLowQual' --filter-name 'SNP_LowQual' --filter-name 'SNP_SOR' -cluster 3 -window 10”

Processing external PTA data
sra files from cord blood tissue with study accession code “SRP178894” were downloaded from the Sequence Read Archive and extracted into bam files using the prefetch and sam-dump tools of the sratoolkit (v2.9.2)[59]. Samtools view (v1.3) was then used with the “-bf 1” argument to select for the paired reads and Picard SamToFastq (v2.24.1) was used with the “RG_TAG=ID” and “OUTPUT_PER_RG=true” arguments to generate fastq files[57,60]. Seqkit replace (v2.2.0) was used to add a sample id to each read name, because they only consisted of a single read number and a number indicating whether it is the first or second read in the pair[61]. Read alignment and variant calling were then performed as described above. Sample names were slightly modified for brevity and to fit our bioinformatic pipeline. The first number in the sample names indicates the treatment concentration and the second number indicates the cell number. PTATO was applied using the random forests trained on the in-house data.

PTATO Nextflow implementation
PTATO was implemented in nextflow (v21.10.6.5661). Submodules are containerized and automatically downloaded by a container engine, allowing for an easy installation. Singularity (v3.8.7-1.el7) was used for this manuscript, though Docker will also work with a small change to the config. As input the pipeline needs a bam file and a vcf containing all variants, both germline and somatic. Additionally, a pre-trained random forest model is necessary when a user wishes to filter somatic variants without training a model. There are also several optional arguments, that allow a user to skip steps by supplying an intermediate file from a previous run or an external file.

PTATO resources
Next to the sample specific inputs, several general resource files were also used to run PTATO, which are listed in PTATO’s “resources.config” file. In order to make PTATO easy to install and more reproducible, these resource files are included with downloads of PTATO. First, the fasta file and accompanying indexes of the hg38 version of the human reference genome were downloaded from GATK (https://gatk.broadinstitute.org/hc/en-us/articles/360035890811). The input files necessary for the COBALT, GRIDSS2, and GRIPSS tools were downloaded from the Hartwig Medical Foundation (https://nextcloud.hartwigmedicalfoundation.nl/s/LTiKTd8XxBqwaiC?path=%2FHMFTools-Resources)[62,63]. A text file containing the centromere locations was downloaded from the UCSC (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=1424951119_QTS0nx5NshNSyspI7KDoJbVh9tci&clade=mammal&org=Human&db=hg38&hgta_group=map&hgta_track=centromeres&hgta_table=0&hgta_regionType=genome&position=chrX%3A15%2C560%2C138-15%2C602%2C945&hgta_outputType=primaryTable&hgta_outFileName=)[64]. A text file with the genomic coordinates of cytobands was also downloaded from the UCSC (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=1424951119_QTS0nx5NshNSyspI7KDoJbVh9tci&clade=mammal&org=Human&db=hg38&hgta_group=map&hgta_track=cytoBand&hgta_table=0&hgta_regionType=genome&position=chrX%3A15%2C560%2C138-15%2C602%2C945&hgta_outputType=primaryTable&hgta_outFileName=). A bed file with the genomic coordinates of simple repeats was downloaded from the UCSC for hg19 (http://genome.ucsc.edu/cgi-bin/hgTables?db=hg19&hgta_group=rep&hgta_track=simpleRepeat&hgta_table=simpleRepeat). A bed file with the genomic coordinates of gene bodies was downloaded from Ensembl for hg19[65]. A bed file with replication timing data was generated as described previously (24). Files for which hg19 versions were downloaded were converted to hg38 using UCSCs LiftOver tool[64]. Shapeit maps for hg38 were included with Shapeit (v4.2.2)[66]. Shapeit reference haplotype vcf files were downloaded from the 1000 genomes project (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/

working/20201028_3202_phased/). Bcftools (v1.9), awk (v4.0.2) and bgzip (v1.8) were used to remove the "chr" prefix from the chromosome names. Our in-house SMuRF tool was used with two in-house mutation blacklists. These are bundled in with PTATO, similar to the other resource files, but they can also be downloaded separately (https://github.com/ToolsVanBox/Genotoxin_assay/tree/main/Blacklist)

QC

Aligment summary metrics were generated for each sample using the CollectAlignmentSummaryMetrics tool from GATK (v4.1.3.0), while WGS metric files were generated using GATKs CollectWGSMetrics tool. Both tools were run using standard parameters. Next, the output of both tools was merged between all the samples and between the tools using R (v4.1.2)[67]. Finally, the R ggplot2 (v3.3.6) package was used to generate quality control figures which are combined in a single pdf[68]. A table containing the merged data used to generate the figures is also written.

Pre-filter somatic variants

Our in-house tool SMuRF (scripts available at: https://github.com/ToolsVanBox/SMuRF) was used to remove germline and low quality variants by applying several filters as described previously[24]. Briefly, we considered variants at autosomal or X chromosomes without any evidence from a paired bulk control sample from the same individual; passed by VariantFiltration with a GATK phred-scaled quality score ≥ 100; a base coverage of at least 10X (30X samples) or 5X (15X samples) in the PTA and paired control sample; a mapping quality (MQ) score of 60; no overlap with single nucleotide polymorphisms (SNPs) in the Single Nucleotide Polymorphism Database v146[69]; and absence of the variant in a panel of unmatched normal human genomes. We additionally filtered base substitutions with a GATK genotype score (GQ) lower than 99 or 10 in PTA or paired control sample, respectively. For indels, we filtered variants with a GQ score lower than 99 in both PTA and paired control sample. In addition, for both SBSs and INDELs, we only considered variants with a variant allele frequency of 0.15 or higher to exclude sequencing artifacts. Lastly, we filtered out mutations that had a VAF of more than 0.3 and/or failed QC in all, or all but one sample in that patient, as this suggests germline mutations that are missed in one or multiple cells due to low quality mapping or low coverage.

Phase germline

Each chromosome was phased separately using Shapeit (v4.2.2), with the raw vcf containing all variants as its input[66]. Additionally, the "sequencing" argument was used, Shapeit maps for the relevant reference genome were supplied to the map argument and a vcf with reference haplotypes was supplied to the reference argument

Allelic imbalance

For each candidate somatic variant, the allelic imbalance was determined using R. For each somatic variant, all phased (germline) variants within 200,000 bp are loaded. To ensure only heterozygous germline variants are used, all variants that are not heterozygous in the bulk sample or do not have a dbSNP reference number were removed. The next steps were done separately for each sample in which the candidate somatic mutation was present. After removing all germline variants that were not heterozygous in the sample, the allele depths of all variants phased to the second allele were swapped and the b-allele frequencies were calculated. Next, the b-allele frequencies were fitted with a locally weighted least squares regression, which was used to predict the b-allele frequency of the candidate somatic variant. This regression was performed using the loess R function with a degree of 2 and using the total allele depth of each variant as weights. Next, a binomial test was performed using both the predicted and observed b-allele frequency as well as the total allele depth of the candidate variant, to determine whether the observed allele frequency of the candidate variant matched the surrounding germline variants. The log of the *p*-value from the allelic imbalance was then used for subsequent steps.

Determine context features

R was used to get the sequence context of each potential mutation of the pre-filtered vcf and functions modified from the R package MutationalPatterns were used to get the mutation type[70]. A bed file was then extracted from the vcf and sorted using bedtools (v2.30.0) with the "sort" argument[71].

To identify the closest genebody and simplerepeat region for each somatic mutation, bedtools was used with the "closest" argument. Since a mutation can sometimes be linked to multiple features from a single bed file, bedtools was then used with the "merge -d -1 -o min" arguments to ensure that each mutation is linked to only the nearest feature for each feature list. To identify the transcriptional strand bias and replication timing for each somatic mutation, bedtools was used with the "intersect" argument. Some mutations were linked to multiple overlapping gene annotations. For the transcriptional stand bias this was solved by using bedtools with the "merge -d -1 -o distinct" arguments to check if a variant was present in the plus strand, minus strand or both. For the replication timing bedtools was used with the "merge -d -1 -o median" arguments to merge mutations that are present in multiple genes. Next, to merge the genebody, simplerepeat, transcriptional strand bias, and replication timing features, bedtools was used with the "intersect" argument, after which the variants were merged using bedtools with the "merge -d -1 -o unique" arguments.

Read-backed phasing
Read-backed phasing is performed using python (v3.6.1). For each heterozygous candidate somatic variant, all overlapping reads are extracted from the sample's bam file. Additionally, all heterozygous germline variants within the area spanned by the reads are extracted from the original input vcf. Next, for each germline variant each read that spans both the germline and somatic variant is checked. Each read that contains either the alternative alleles for both the germline and somatic variant or the reference alleles for both the germline and somatic variant is counted as a cis read. Other reads are counted as trans reads. If a candidate is real, then it would be expected that almost all reads are either cis or trans. Whether the variants are cis, trans, or mixed is then calculated based on a Bayesian likelihood score similar to the one used by SVTyper[72]. The likelihood scores of the three options are then combined into a single Phred-scaled quality score.

Random forest training data
A true positive and an artifact dataset were created to train the random forest model. For the true positive set, we included candidate somatic mutations that were present in multiple samples from a single donor, as these are likely true mutations that originated in an ancestor cell. For the artifact set we included candidate somatic mutations that had a read-backed phasing score below 1. Any variants that were shared and also had a read-backed phasing score of less than 1 were excluded from both the true positive and the artifact datasets. The IBFM26, IBFM35 and PMCAHH1-FANCKO samples contained several copy-number changes and loss of heterozygosity sites. Variants within these sites were also excluded from training.

Additionally, for the SBSs, any variant in the cord blood samples that was not shared with another sample was also considered an artifact, as the number of true mutations in the cord bloods is expected to be very low[18]. Finally, we subsampled the number of artifactual SBSs to be the same as the number of true SBSs, to result in a better class balance. These last two steps were not done for the indels, as we already had a large collection of artifactual indels and the class imbalance in the indels reflects the real imbalance seen in PTA data.

Random forest
A random forest was trained on the previously described features with the randomForest (v 4.7-1) R package supplying the "mtry" argument with a value of 4. For some variants no *p*-value for the allelic imbalance or no replication timing value could be calculated, therefore they were excluded from the training. Instead, two more random forests were trained that did include these variables. One without the allelic imbalance variable and one without both this variable and the replication timing variable.

When using the random forest model to filter candidate somatic mutations, a prediction score is calculated for each variant per sample. Next, any variants with either a low (<1) or a high (>=1000) read-backed phasing score are used as artifact and true positive mutation sets to validate the performance of the random forest on the sample it is filtering. The precision and recall of the model is calculated for different cutoffs of the prediction score generated by the random forest. This is done for cutoffs between 0 and 1 with steps of 0.01. The intersection of the precision and recall curves is then used as the final cutoff for the random forest model. Next, any candidate somatic mutation that is predicted to be an artifact by the random forest or that has a low read-backed phasing score is filtered out per sample. A vcf containing all mutations annotated with the random forest score is also written out as is a table containing the precision and recall at different cutoffs, allowing users of the model to modify the stringency of the mutation filtering when needed. Finally, a multi-sample vcf is written containing the potential somatic variants of all the samples in a donor and the associated random forest prediction scores.

PTATO validation
The performance of the walker and random forests on the training data and the external cord blood data was analyzed using R (4.1.0). For the training data the out-of-bag predictions were used. The mutational patterns and signature analyses were made using MutationalPatterns (v3.7.1)[70]. Mutational signatures were used from COSMIC (v3.2) as well as the previously described HSPC, PTA, and ENU signatures[18,33,36,73]. Figures were made using ggplot2 (v3.3.5)[68].

The HSPC and PTA signatures were used for signature refitting on the validation data, because the HSPC signature is known to be the main signature in adult blood cells[18]. In contrast, the SBS5 signature was used with the external cord blood cells, as cord blood cells are known to not yet contain a contribution of the HSPC signature[74]. In agreement with this, the cosine similarity of the reconstructed mutation profile with the original was reduced when the SBS5 signature was used with the validation data and when the HSPC signature was used with the external cord blood cells.

To match the number of mutations in the HSC and MPP of the AML patient to HSPCs of healthy donors with different ages, we selected the autosomal variants and extrapolated the mutation load to the entire called autosomal genome based on the surveyed fraction of the genome, as previously described[24]. Next, we used a linear mixed-effects model to fit the extrapolated mutation load to the donor age for the HSPCs. 95% confidence and 95% prediction intervals were calculated using the R package ggeffects (v1.1.0)[75]. The extrapolated mutation load of the HSC and MPP of the AML patient were then compared to the fitted distribution of the HSPCs.

Shared mutation heatmap and lineage tree creation

A matrix was created with all autosomal mutations that were clonally present in at least two samples of one AML donor and that passed PTATO in at least one sample. A heatmap was then plotted showing per variant per sample if a variant was clonally present, absent or had failed the quality control.

For the lineage tree a matrix was made that marked for each autosomal variant for each sample if the variant was either present or not based on the genotype. Variants had to pass PTATO in at least one sample, but they did not have to pass the quality control or be clonally present in other samples as this causes too many mutations to be missed. A phylogenetic tree was then generated using neighbor-joining as described previously[74].

Timing of driver occurrence

The drivers in the AML patient were timed using two different orthogonal methods. First, we matched the number of autosomal mutations in the branches containing drivers to HSPCs of healthy donors with different ages, similar to the comparison we described above. However, instead of using the mutation load directly, we fit their somatic mutations to the SBS1, SBS5, SBS18, and HSPC signatures. We then removed the contribution of the SBS18 signature as this signature's contribution is not correlated with age and might have an increased contribution caused by the drivers. Next, we extrapolated the combined contribution of the remaining signatures to the entire callable autosomal region of the genome as above. For the branches containing drivers the average surveyed fraction of the genome of the cells within the relevant clade was used. Next, we used a linear mixed-effects model to fit the extrapolated mutation load of the HSPCs to the donor age. The extrapolated mutation loads of branches containing drivers were then compared to the fitted distribution of the HSPCs, to estimate the age of occurrence.

The second method to time the occurrence of the drivers was to divide the number of autosomal somatic mutations in a branch containing the driver of interest by the average of the total number of mutations in the cells carrying this driver and then multiplying this number by the age of the patient. For the timing of the *NUP98-NSD1* translocation and *IDH2* mutation we used cells that did not also contain the *WT1* and *FLT3* drivers, as these might have changed the mutation rate.

Normalization of copy number ratios for SV detection

GC-normalized read depth per 1000 basepair genomic window was calculated by COBALT (v1.11). A coverage panel of normals (PON) was generated by merging COBALT ratio files of 12 copy number neutral PTA-based samples. The total read counts from all windows of each sample were first normalized so that every sample has the same total amount of read counts. Subsequently the mean readcount per bin over all normal samples in the PON was calculated. PTATO uses the coverage PON file to smoothen PTA-specific coverage fluctuations. First the total read depth in a test sample is normalized to the same total amount of read counts in the coverage PON. Subsequently the read counts in each window are divided by the mean read counts in the same window in the PON. Additionally, the bottom and top 1% outlier windows in the PON file and the windows located within 1Mb distance of centromeres and telomers are excluded from the analysis.

The smoothened read counts were subsequently binned in 100kb windows. The copynumber (v1.34.0) R-package with parameter "gamma=100" was used to segment the median read count data in both the 100kb and 1kb windows[76]. The segments based on the 100kb resolution were used as raw copy number segments. The start and end coordinates of these raw copy number segments were fine mapped by taking the start and end coordinates of overlapping 1kb window-based segments. Fine mapped segments with a copy number ratio of <1.5 were considered to be copy number losses and segments with ratios >2.5 were considered to be copy number gains.

Deviation of allele frequency calculations

The VAFs for each germline SNV were collected from the corresponding bulk control sample. To reduce noise due to uneven amplification, germline SNVs were binned in 100kb windows instead of taking B-allele frequencies of each individual variant. To calculate a mean allele frequency for multiple variants in a bin, we calculated the deviation of allele frequency (DAF) by taking the absolute value after subtracting the VAF of each variant from 0.5, which is the expected VAF for a perfectly amplified and sequenced germline variant. Thus, each variant has a DAF between 0 (corresponding to a VAF of 0.5) and 0.5 (corresponding to a VAF of 0 or 1). Subsequently all DAF values are binned in 100kb genomic regions and the mean DAF for each bin is calculated. The copynumber R-package with parameter "gamma=100" was used to segment the 100kb bins in crude DAF regions. These crude segments were fine mapped by adjusting the start and end coordinates of the segments to the positions of the nearest germline SNVs, that were within 200kb of the segment, with similar DAFs as the segment. Segments with a DAF of more than 0.4 (corresponding to VAF

< 0.1 or > 0.9) were considered to be loss of heterozygosity regions. Segments with a DAF between 0.16 and 0.4 (corresponding to VAFs between 0.1 to 0.32 or 0.64 to 0.9) were considered to be regions with copy number gains.

SV breakend calling and filtering

Somatic SV breakends were called by GRIDSS v2.13.2 and prefiltered by GRIPSS v1.9 using a corresponding bulk-sequenced germline control[63,77]. The GRIPSS-filtered somatic breakends of 15 PTA-based samples of four unrelated individuals were merged using bedtools merge (v2.30.0). Breakend positions occurring within 2000bp of each other in multiple of these individuals were included in a breakend PON. Candidate breakends in other samples overlapping with the regions in the breakend PON were filtered. Subsequently the normalized coverage and DAF of the SV candidates was calculated. Breakends of duplications were filtered if the DAF was less than 0.18 and/or the copy number ratio was <2.5. Breakends of deletions were filtered if the DAF was less than 0.4 and/or the copy number ratio was >1.5. Breakends with a coverage of more than 100 were also excluded for samples with a targeted genome coverage of 15x as many artefacts occur in these regions with excess coverage. Inversions were filtered if they only have one breakpoint junction. Additionally, all inversions less than 1kb in size were filtered. Interchromosomal events were also filtered if they only have one breakpoint junction, unless they were situated less than 100kb from a copy number variant. This exception rescues unbalanced translocations.

Integration of coverage, allele frequencies and structural variant breakends

The coverage segments, DAF segments, and breakends of SV candidates were intersected to create the final list of filtered structural variants. Copy number changes were required to have both coverage and DAF support, but not necessarily breakend support, as many CNVs have start and/or end positions within repeat regions that are difficult to capture with PTA and/or short-read sequencing. Regions with a DAF >0.4 (corresponding to VAFs of <0.1 and >0.9) without coverage support (copy number >1.5) were considered to be loss-of-heterozygosity regions. ggplot2 and Circos (v0.69-9) were used for to visualize structural variants and karyograms[78].

Author Contributions

FM, SM, and RvB jointly wrote the manuscript. FM, SM, and MJvR wrote PTATO. FM, SM, MJvR, and JdK performed bioinformatic analyses. SM, EB, IvdW, EA, DR, NMG, MV, and AMB generated the samples. RvB supervised the study.

Acknowledgments

The authors would like to thank the Hartwig Medical Foundation (Amsterdam, the Netherlands) for facilitating low-input whole-genome sequencing. We thank the Central Biobank of the UMCU for providing umbilical cord blood samples and the GPOH for providing Fanconi anemia samples. This study was financially supported by a grant from the Dutch Cancer Society (KWF Research Project 12682) to R. van Boxtel.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Availability of data and materials

PTATO is available at: https://github.com/ToolsVanBox/PTATO/tree/main
The datasets supporting this article and the scripts that can be used to reproduce all the figures in this paper are available upon request.

References

1. Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, et al. The mutational landscape of human somatic and germline cells. Nature. 2021;597(7876):381–6.
2. Manders F, van Boxtel R, Middelkamp S. The Dynamics of Somatic Mutagenesis During Life in Humans. Vol. 2, Frontiers in Aging. 2021.
3. Tang X, Huang Y, Lei J, Luo H, Zhu X. The single-cell sequencing: new developments and medical applications. Cell Biosci. 2019;9(1):53.
4. Chongyi C, Dong X, Longzhi T, Heng L, Guangyu Z, Lei H, et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science (80-). 2017 Apr 14;356(6334):189–94.
5. Veronica G-P, Sivaraman N, Yuntao X, David K, Robert C, Yakun P, et al. Accurate genomic variant detection in single cells with primary template-directed amplification. Proc Natl Acad Sci. 2021 Jun 15;118(24):e2024176118.
6. Dou Y, Gold HD, Luquette LJ, Park PJ. Detecting Somatic Mutations in Normal Cells. Trends Genet. 2018;34(7):545–57.
7. Desai P, Mencia-Trinchant N, Savenkov O, Simon MS, Cheang G, Lee S, et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. Nat Med. 2018;24(7):1015–23.
8. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Yu K, Tarabichi M, et al. The evolutionary history of 2 ,658 cancers. BioRxiv. 2017;578(August 2017).
9. Mitchell TJ, Turajlic S, Rowan A, Nicol D, Farmery JHR, O'Brien T, et al. Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. Cell. 2018;173(3):611-623.e17.
10. Lee JJ-K, Park S, Park H, Kim S, Lee J, Lee J, et al. Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma. Cell. 2019;177(7):1842-1857.e21.

11. Rustad EH, Yellapantula V, Leongamornlert D, Bolli N, Ledergor G, Nadeu F, et al. Timing the initiation of multiple myeloma. Nat Commun. 2020;11(1):1917.
12. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018;15(2):81–94.
13. Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature. 2010;467(7319):1109–13.
14. Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature. 2014;513(7518):422–5.
15. Park S, Mali NM, Kim R, Choi J-W, Lee J, Lim J, et al. Clonal dynamics in early human embryogenesis inferred from somatic mutation. Nature. 2021;597(7876):393–7.
16. Coorens THH, Moore L, Robinson PS, Sanghvi R, Christopher J, Hewinson J, et al. Extensive phylogenies of human development inferred from somatic mutations. Nature. 2021;597(7876):387–92.
17. Roerink SF, Sasaki N, Lee-Six H, Young MD, Alexandrov LB, Behjati S, et al. Intra-tumour diversification in colorectal cancer at the single-cell level. Nature [Internet]. 2018;556(7702):457–62. Available from: https://doi.org/10.1038/s41586-018-0024-3
18. Osorio FG, Rosendahl Huber A, Oka R, Verheul M, Patel SH, Hasaart K, et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. 2018 Nov;25(9):2308-2316.e4.
19. Brandsma AM, Bertrums EJM, van Roosmalen MJ, Hofman DA, Oka R, Verheul M, et al. Mutation Signatures of Pediatric Acute Myeloid Leukemia and Normal Blood Progenitors Associated with Differential Patient Outcomes. Blood Cancer Discov. 2021 Sep 1;2(5):484 LP – 499.
20. Yoshida K, Gowers KHC, Lee-Six H, Chandrasekharan DP, Coorens T, Maughan EF, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature. 2020;578(7794):266–72.
21. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;574(7779):532–7.
22. Wen L, Tang F. Recent advances in single-cell sequencing technologies. Precis Clin Med. 2022 Mar 22;5(1):pbac002.
23. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC, et al. The origin and evolution of mutations in acute myeloid leukemia. Cell. 2012;150(2):264–78.
24. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538(7624):260–4.
25. Franco I, Johansson A, Olsson K, Vrta⍰nik P, Lundin P, Helgadottir HT, et al. Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. Nat Commun. 2018;9(1):800.
26. Lee-six H, Øbro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature. 2018;
27. Mitchell E, Spencer Chapman M, Williams N, Dawson KJ, Mende N, Calderbank EF, et al. Clonal dynamics of haematopoiesis across the human lifespan. Nature. 2022;606(7913):343–50.
28. Ellis P, Moore L, Sanders MA, Butler TM, Brunner SF, Lee-Six H, et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. Nat Protoc. 2021;16(2):841–71.
29. Grossmann S, Hooks Y, Wilson L, Moore L, O'Neill L, Martincorena I, et al. Development, maturation, and maintenance of human prostate inferred from somatic mutations. Cell Stem Cell. 2021;1–13.
30. Dean FB, Nelson JR, Giesler TL, Lasken RS. Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. Genome Res. 2001 Jun 1;11(6):1095–9.
31. B. DF, Seiyu H, Linhua F, Xiaohong W, Fawad FA, Patricia B-W, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci. 2002 Apr 16;99(8):5261–6.
32. Miller MB, Huang AY, Kim J, Zhou Z, Kirkham SL, Maury EA, et al. Somatic genomic changes in single Alzheimer's disease neurons. Nature. 2022;604(7907):714–22.
33. Luquette LJ, Miller MB, Zhou Z, Bohrson CL, Galor A, Lodato MA, et al. Ultraspecific somatic SNV and indel detection in single neurons using primary template-directed amplification. bioRxiv. 2021 Jan 1;2021.04.30.442032.
34. Bohrson CL, Barton AR, Lodato MA, Rodin RE, Luquette LJ, Viswanadham V V, et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nat Genet. 2019;51(4):749–54.
35. Drost J, van Boxtel R, Blokzijl F, Mizutani T, Sasaki N, Sasselli V, et al. Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science (80-). 2017;238:eaao3130.
36. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177:821-836.e16.
37. Mallory XF, Edrisi M, Navin N, Nakhleh L. Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biol. 2020;21(1):208.
38. Dong X, Zhang L, Hao X, Wang T, Vijg J. SCCNV: A Software Tool for Identifying Copy Number Variation From Single-Cell Whole-Genome Sequencing. Vol. 11, Frontiers in Genetics. 2020.
39. Lasken RS, Stockwell TB. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. BMC Biotechnol. 2007;7(1):19.
40. Reikvam H, Hatfield KJ, Kittang AO, Hovland R, Bruserud Ø. Acute Myeloid Leukemia with the t(8;21) Translocation: Clinical Consequences and Biological Implications. Ouhtit A, editor. J Biomed Biotechnol. 2011;2011:104631.
41. Shen X, Wang R, Kim MJ, Hu Q, Hsu CC, Yao J, et al. A Surge of DNA Damage Links Transcriptional Reprogramming and Hematopoietic Deficit in Fanconi Anemia. Mol Cell. 2020;80(6):1013-1024.e6.
42. Morita K, Wang F, Jahn K, Hu T, Tanaka T, Sasaki Y, et al. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. Nat Commun. 2020;11(1):5327.
43. Ediriwickrema A, Aleshin A, Reiter JG, Corces MR, Köhnke T, Stafford M, et al. Single-cell mutational profiling enhances the clinical evaluation of AML MRD. Blood Adv. 2020 Mar 9;4(5):943–52.
44. Miles LA, Bowman RL, Merlinsky TR, Csete IS, Ooi AT, Durruthy-Durruthy R, et al. Single-cell mutation analysis of clonal evolution in myeloid malignancies. Nature. 2020;587(7834):477–82.
45. Romer-Seibert JS, Meyer SE. Genetic heterogeneity and clonal evolution in acute myeloid leukemia. Curr Opin Hematol. 2021;28(1).
46. de Kanter JK, Peci F, Bertrums E, Rosendahl Huber A, van Leeuwen A, van Roosmalen MJ, et al. Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. Cell Stem Cell. 2021;
47. Anupriya A, J. BW, B. WD, A. EC, B. OS, Guang F, et al. Differentiation of leukemic blasts is not completely blocked in acute myeloid leukemia. Proc Natl Acad Sci. 2019 Dec 3;116(49):24593–9.
48. Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature. 2019;574(7779):538–42.
49. Moore L, Leongamornlert D, Coorens THH, Sanders MA, Ellis P, Dentro SC, et al. The mutational landscape of normal human endometrial epithelium. Nature. 2020;580(7805):640–6.
50. Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing S V, Ellis P, et al. Somatic mutation landscapes at single-molecule resolution. Nature. 2021;593(7859):405–10.
51. Dong X, Longzhi T, Chi-Han C, Heng L, Sunney XX. Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. Proc Natl Acad Sci. 2021 Feb 23;118(8):e2013106118.
52. Zong C, Lu S, Chapman AR, Xie XS. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. Science (80-). 2012 Dec 21;338(6114):1622–6.
53. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518:360–4.
54. Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science (80-). 2019 Jun 28;364(6447):eaaw2872.
55. Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. Local Determinants of the Mutational Landscape of the Human Genome. Cell. 2019;177:101–14.

56. Xia Y, Gonzales-Pena V, Klein DJ, Luquette JJ, Puzon L, Siddiqui N, et al. Genome-Wide Disease Screening in Early Human Embryos with Primary Template-Directed Amplification. bioRxiv. 2021 Jan 1;2021.07.06.451077.
57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Jun 8;25(16):2078–9.
58. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.
59. Leinonen R, Sugawara H, Shumway M, Collaboration on behalf of the INSD. The Sequence Read Archive. Nucleic Acids Res. 2011 Jan 1;39(suppl_1):D19–21.
60. Picard toolkit. Broad Institute, GitHub repository. Broad Institute; 2019.
61. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One. 2016 Oct 5;11(10):e0163962.
62. Cameron DL, Baber J, Shale C, Papenfuss AT, Valle-Inclan JE, Besselink N, et al. GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. bioRxiv. 2019 Jan 1;781013.
63. Cameron DL, Baber J, Shale C, Valle-Inclan JE, Besselink N, van Hoeck A, et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. 2021;22(1):202.
64. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002 Jun 1;12(6):996–1006.
65. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;48:D682–8.
66. Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019;10(1):5436.
67. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2018.
68. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016.
69. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308–11.
70. Manders F, Brandsma AM, de Kanter J, Verheul M, Oka R, van Roosmalen MJ, et al. MutationalPatterns: the one stop shop for the analysis of mutational processes. BMC Genomics. 2022;23(1):134.
71. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010/01/28. 2010 Mar 15;26(6):841–2.
72. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12(10):966–8.
73. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
74. Hasaart KAL, Manders F, van der Hoorn M-L, Verheul M, Poplonski T, Kuijk E, et al. Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis. Sci Rep. 2020;10(1):12991.
75. Lüdecke D. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. J Open Source Softw. 2018;3(26):772.
76. Nilsen G, Liestøl K, Van Loo P, Moen Vollan HK, Eide MB, Rueda OM, et al. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics. 2012;13(1):591.
77. Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, et al. GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res. 2017 Dec 1;27(12):2050–60.
78. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res. 2009 Sep 1;19(9):1639–45.

Supplementary material

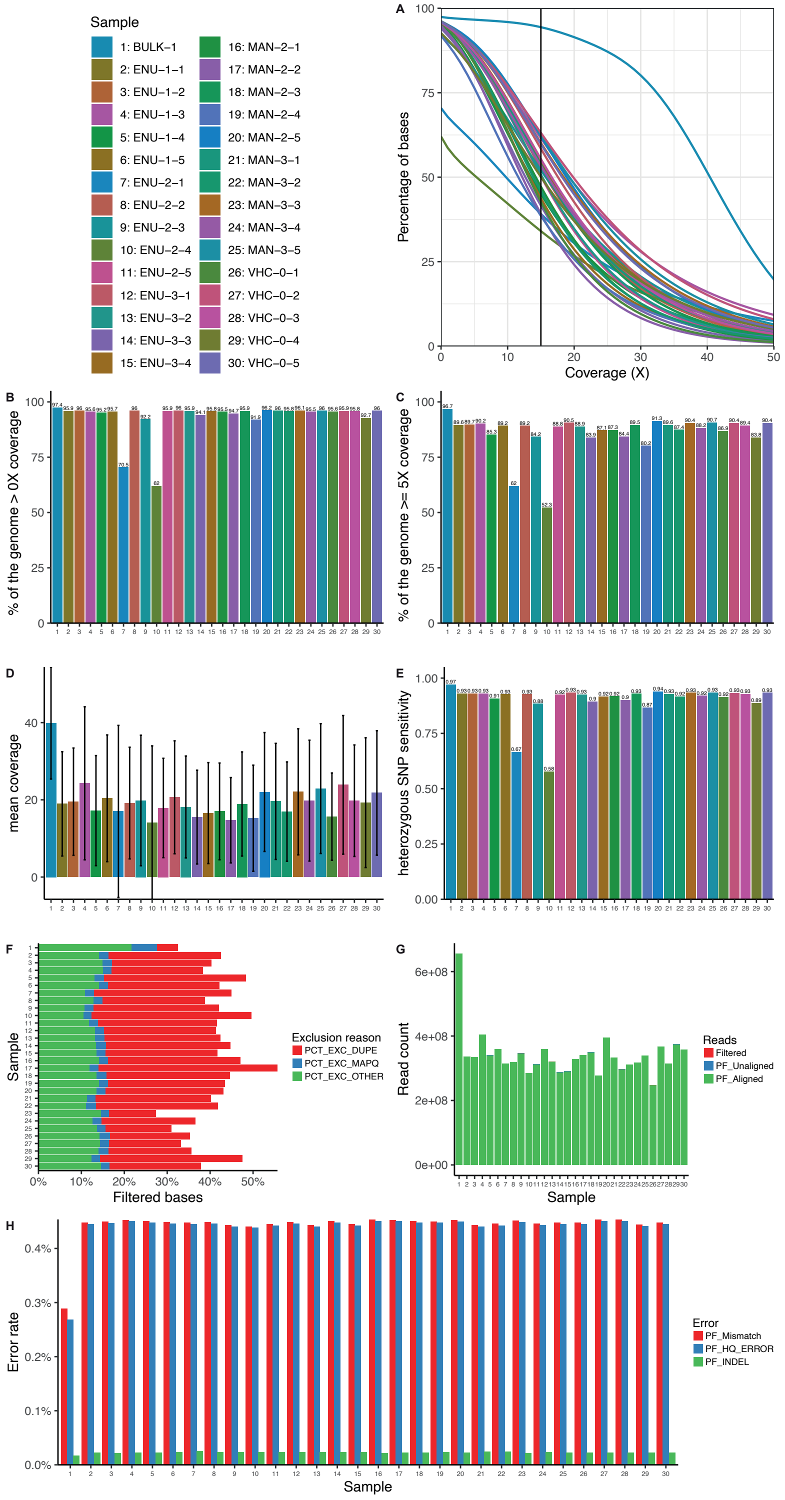


Figure S1: Quality control plots show quality differences between samples.

a The percentage of bases with at least x coverage is plotted against the coverage per sample. **b** Bar plot showing the percentage of the genome with a read coverage of more than 0 per sample. **c** Bar plot showing the percentage of the genome with a read coverage of at least 5 per sample. **d** Bar plot depicting the mean read coverage per sample. The error bars show the standard deviation. **e** Bar plot depicting the heterozygous SNP sensitivity per sample. The heterozygous SNP sensitivity is a theoretical estimate of the sensitivity to detect heterozygous SNPs based on the coverage and base quality distributions. **f** Bar plot depicting the percentage of bases that was filtered per sample. The color indicates the exclusion reason, which can be a base being in a read marked as a duplicate, low mapping quality, or a different reason. The other reasons are further delineated in the quality control table generated by PTATO. **g** Bar plot depicting the number of reads per sample. The color indicates if the read was filtered out, unaligned or aligned. PF stands for reads that have passed Illumina's filters. **h** Bar plot depicting the error rate per sample. This is shown separately for the percentage of mismatched bases in aligned reads, the percentage of mismatched bases in aligned reads with a mapping quality of at least 20, and the number of indels per 100 aligned bases.

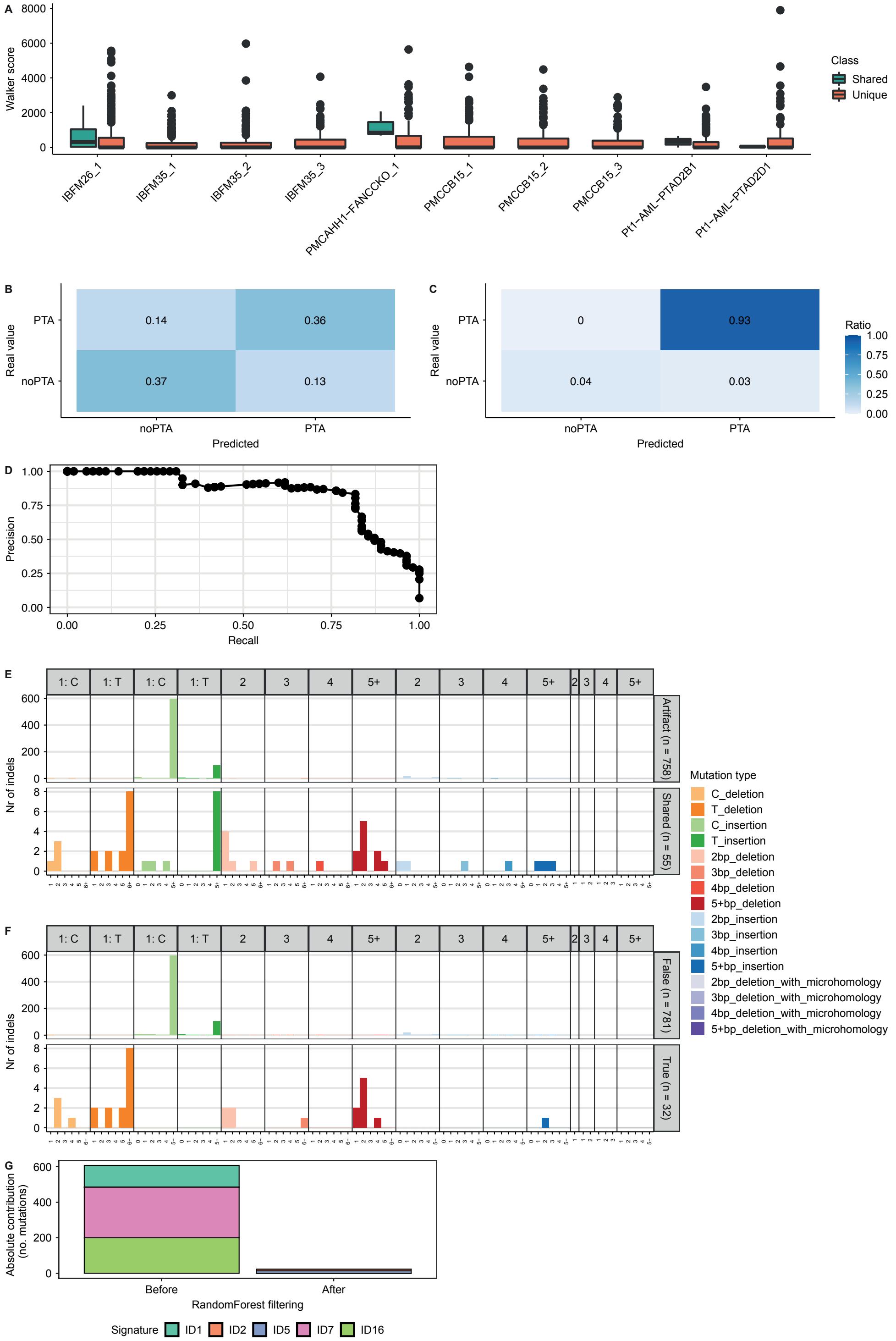


Figure S2: The walker and random forest perform well on the indel training data.
a Boxplot showing the walker scores of the indels in the samples used to train the random forest. The color indicates if mutations are shared between samples or are unique. Unique mutations with a walker score below 0 were used to train the random forest. Not all samples have calculated walker scores for shared indels, because there are a limited number of shared indels and the majority of them are not close enough to a heterozygous germline variant to calculate a walker score. **b** Heatmap depicting the ratio of true positives, false positives, true negatives and false negatives for SBSs. **c** Heatmap depicting the ratio of true positives, false positives, true negatives and false negatives for Indels. **d** Precision and recall curve showing the performance of the random forest using all input variables on the out-of-bag training data for cutoffs between 0 and 1 with a step of 0.01. **e** Absolute contribution of each indicated mutation type to the indel mutation spectrum for the mutations used to train the random forests separated into the artifact variants (samples = 10) and the mutations that were shared between samples (samples = 10). **f** Absolute contribution of each indicated mutation type to the indel mutation spectrum for the mutations predicted to be artifacts (samples = 10) and the mutations predicted to be true somatic mutations (samples = 10). **g** Absolute contribution of each mutational signature for all the mutations used to train the random forests (n = 813; samples = 10) and for all the mutations predicted to be true (n = 32; samples = 10) by the random forests.

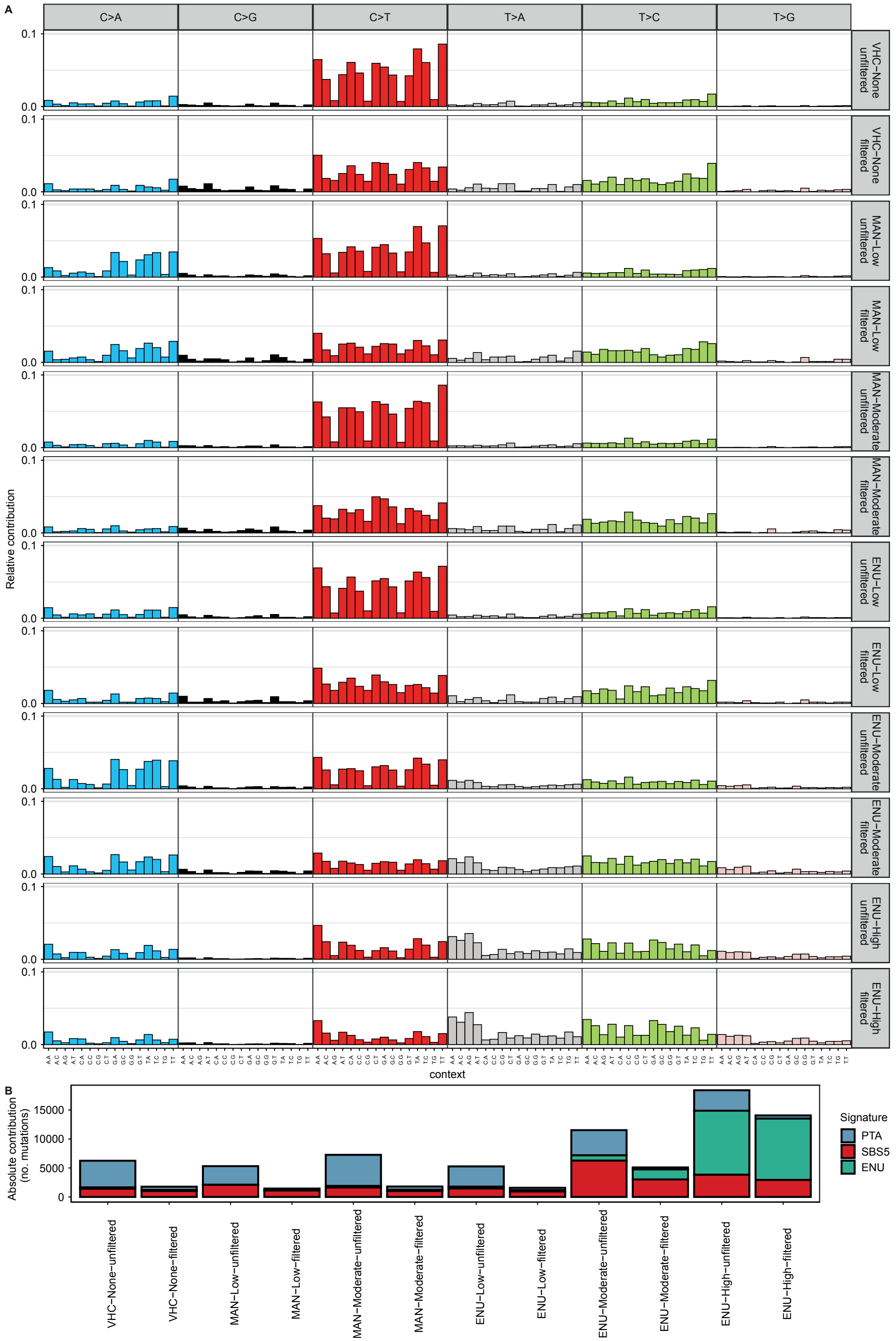
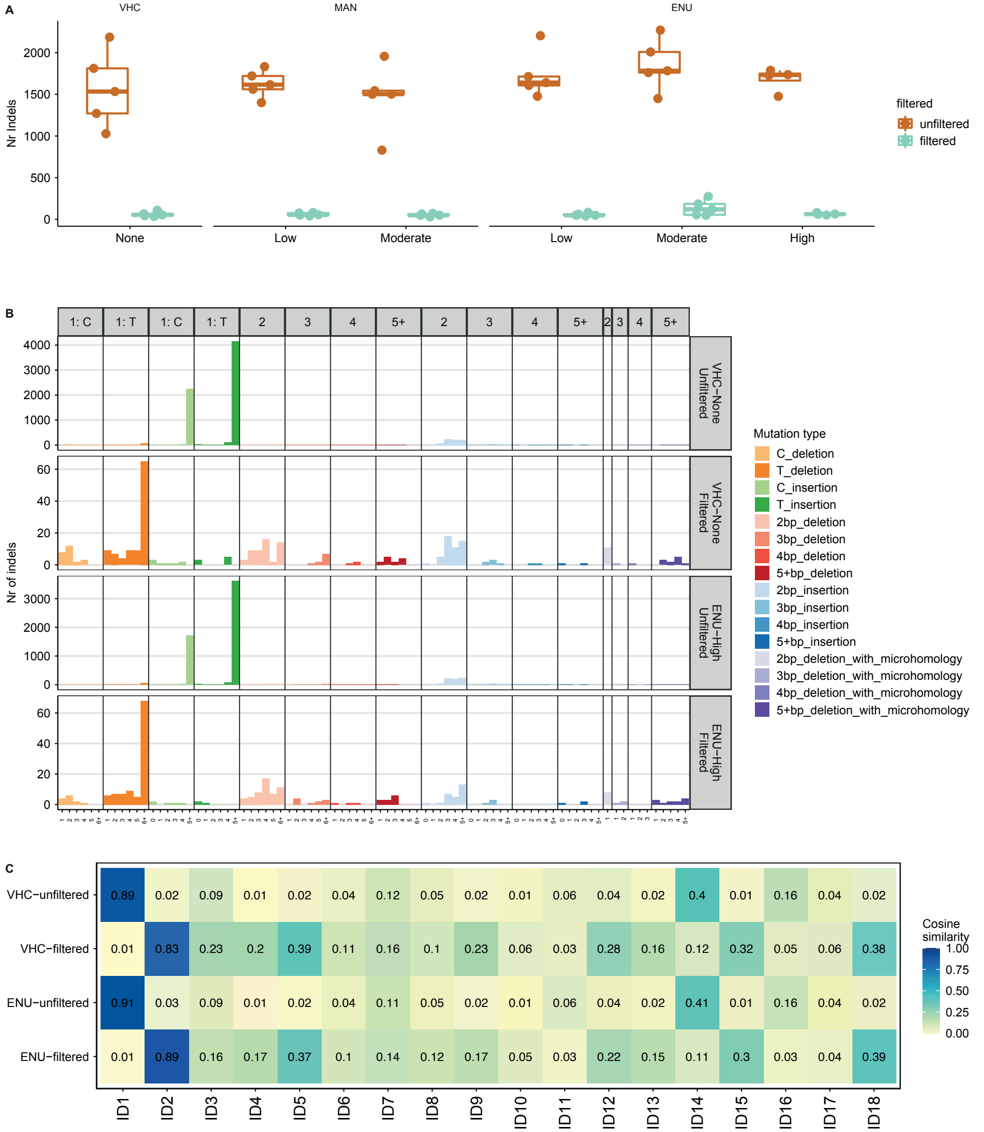


Figure S3: PTATO accurately filters SBSs in external cord blood samples in different treatment categories. **a** Relative contribution of each trinucleotide change to the point mutation spectra of the cord blood samples treated with a vehicle control (unfiltered: n = 5646; filtered: n = 1787; samples = 5), low concentration of MAN (unfiltered: n = 5375; filtered: n = 1624; samples = 5), moderate concentration of MAN (unfiltered: n = 6502; filtered: n = 1811; samples = 5), low concentration of ENU (unfiltered: n = 4839; filtered: n = 1607; samples = 5), moderate concentration of ENU (unfiltered: n = 12943; filtered: n = 6106; samples = 5), or with a high concentration of ENU (unfiltered: n = 21152; filtered: n = 16923; samples = 4) before and after filtering with PTATO. The mutations of samples with the same treatment were pooled together. **b** Absolute contribution of mutational signatures PTA, ENU, and SBS5 to the cord blood samples treated with a vehicle control (unfiltered: n = 5646; filtered: n = 1787; samples = 5), low concentration of MAN (unfiltered: n = 5375; filtered: n = 1624; samples = 5), moderate concentration of MAN (unfiltered: n = 6502; filtered: n = 1811; samples = 5), low concentration of ENU (unfiltered: n = 4839; filtered: n = 1607; samples = 5), moderate concentration of ENU (unfiltered: n = 12943; filtered: n = 6106; samples = 5), or with a high concentration of ENU (unfiltered: n = 21152; filtered: n = 16923; samples = 4) before and after filtering with PTATO. The mutations of samples with the same treatment were pooled together.



	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9	ID10	ID11	ID12	ID13	ID14	ID15	ID16	ID17	ID18
VHC−unfiltered	0.89	0.02	0.09	0.01	0.02	0.04	0.12	0.05	0.02	0.01	0.06	0.04	0.02	0.4	0.01	0.16	0.04	0.02
VHC−filtered	0.01	0.83	0.23	0.2	0.39	0.11	0.16	0.1	0.23	0.06	0.03	0.28	0.16	0.12	0.32	0.05	0.06	0.38
ENU−unfiltered	0.91	0.03	0.09	0.01	0.02	0.04	0.11	0.05	0.02	0.01	0.06	0.04	0.02	0.41	0.01	0.16	0.04	0.02
ENU−filtered	0.01	0.89	0.16	0.17	0.37	0.1	0.14	0.12	0.17	0.05	0.03	0.22	0.15	0.11	0.3	0.03	0.04	0.39

Figure S4: PTATO accurately filters indels in external cord blood samples.
a Boxplot showing the number of indels for human cord blood samples treated with different concentrations of a vehicle control (VHC; n = 5), D-mannitol (MAN; low: n = 5, moderate: n = 5) or N-ethyl-N-nitrosourea (ENU; low: n = 5, moderate: n = 5, high: n = 4). The color indicates if the mutations were filtered by PTATO. **b** Absolute contribution of each indicated mutation type to the indel mutation spectrum of the cord blood samples treated with a vehicle control (unfiltered: n = 8506; filtered: n = 4199; samples = 5) or with a high concentration of ENU (unfiltered: n = 7335; filtered: n = 3562; samples = 4) before and after filtering with PTATO. The mutations of samples with the same treatment were pooled together. **c** Heatmap showing the cosine similarity between the COSMIC indel signatures and the indel mutation spectra of the cord blood samples treated with a vehicle control (unfiltered: n = 8506; filtered: n = 4199; samples = 5) or with a high concentration of ENU (unfiltered: n = 7335; filtered: n = 3562; samples = 4) before and after filtering with PTATO.

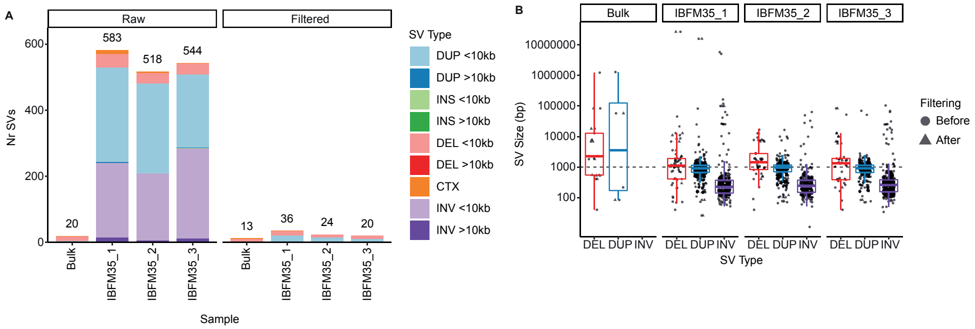


Figure S5: PTATO filters out many SVs in HSPC samples from a donor with Fanconi anemia.
a The number of SVs is shown before and after filtering with PTATO for a bulk sequenced AML sample and three PTA-based HSPC samples from a donor with Fanconi anemia. The type of structural variant is indicated by the color. **b** The sizes of the deletions, duplications, and inversions in a bulk sequenced AML sample and three PTA-based HSPC samples from a donor with Fanconi anemia are shown before and after filtering. The shape of a datapoint indicates whether the size is before or after filtering. DUP = Duplication; INS = Insertion; DEL = Deletion; CTX = Inter-chromosomal translocation; INV = Inversion.

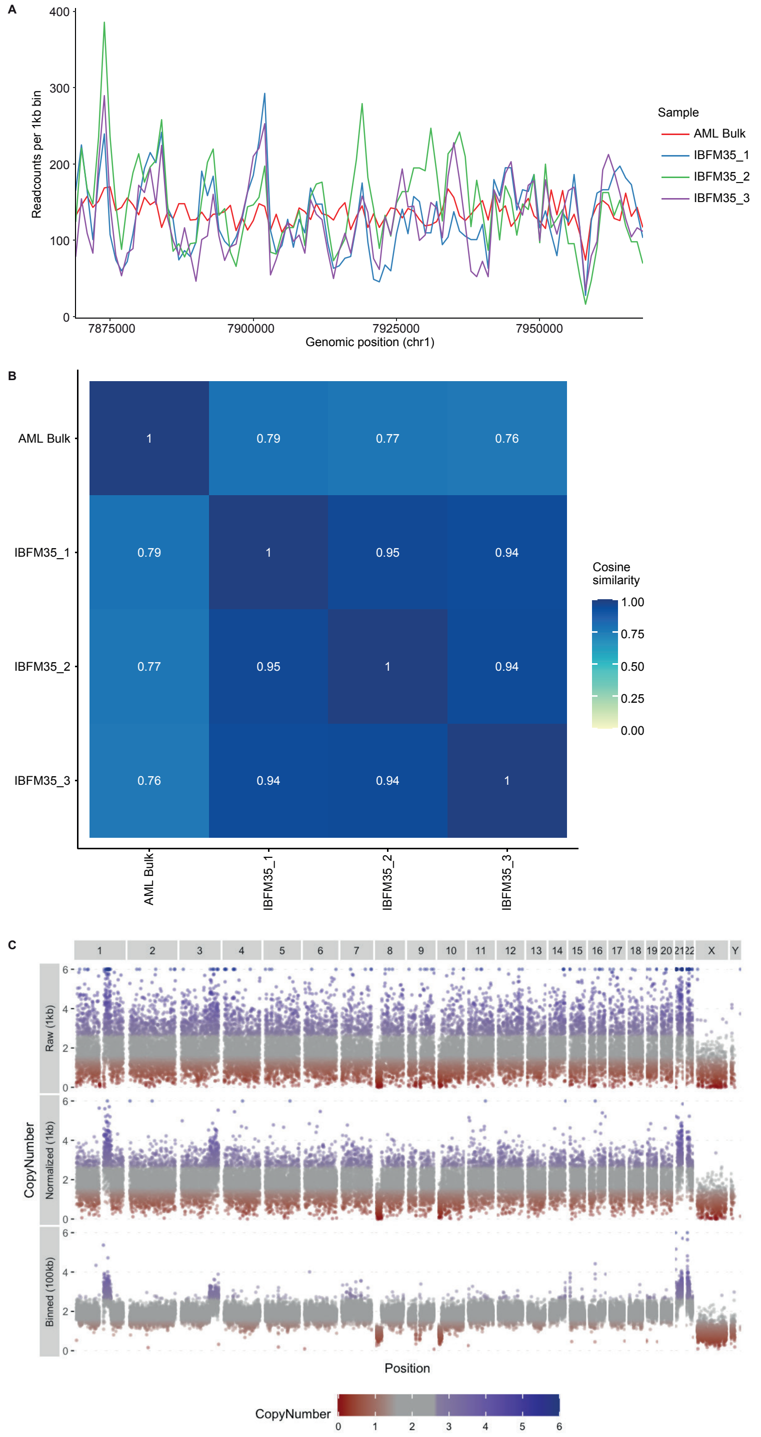


Figure S6: PTATO normalizes and bins the read-depths of PTA-based data.
a The number of reads per 1kb bin in a bulk sequenced AML sample and three PTA-based HSPC samples

from a donor with Fanconi anemia is shown for a 100kb fraction of chromosome 1. **b** Heatmap depicting the cosine similarities of read counts in 1kb bins between a bulk sequenced AML sample and three PTA-based HSPC samples from a donor with Fanconi anemia. **c** The copy numbers calculated by COBALT are shown for sample IBFM35_2 across the autosomal and sex chromosomes. Copy numbers are shown based on the number of reads per 1kb bin (top). Additionally, the copy numbers are shown after the 1kb bins have been normalized for the recurrence in PTA data (middle), and after the normalized bins have been binned in larger 100kb bins (bottom). The color indicates the estimated copy number.

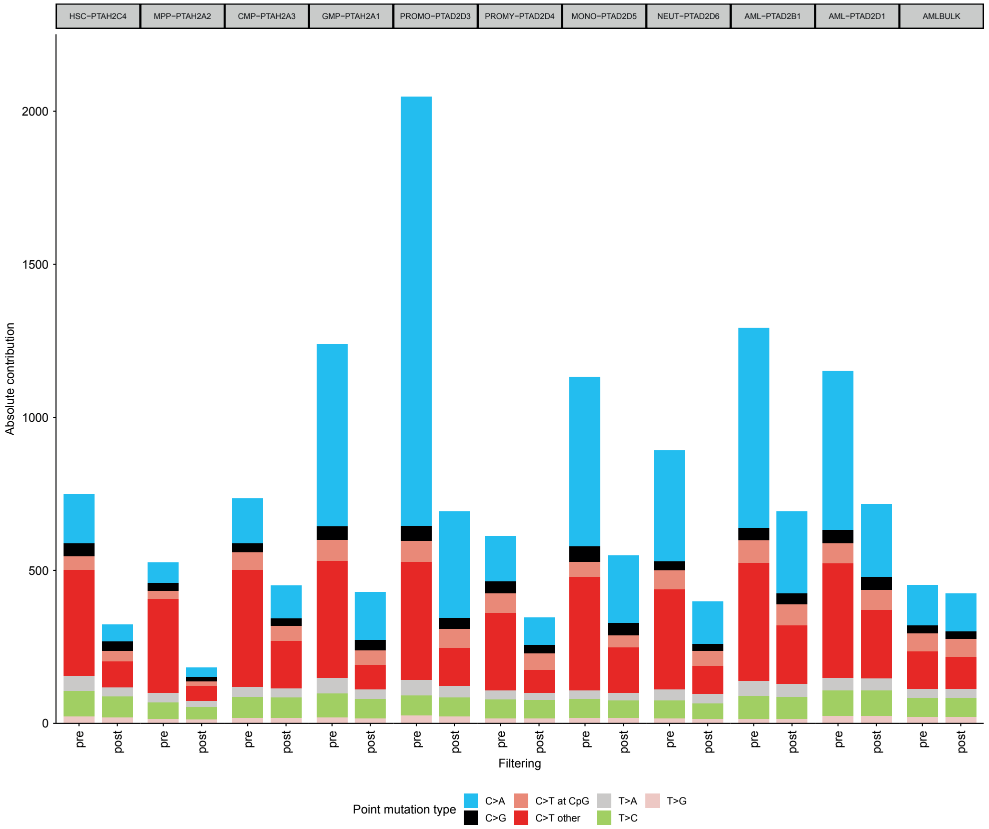


Figure S7: PTATO filters out many artifacts in both differentiated and stem cells.
Spectrum of the six types of base substitutions for different cell types in an AML patient. The colors indicate the type of base substitution. The number of base substitutions is shown before and after filtering with PTATO.

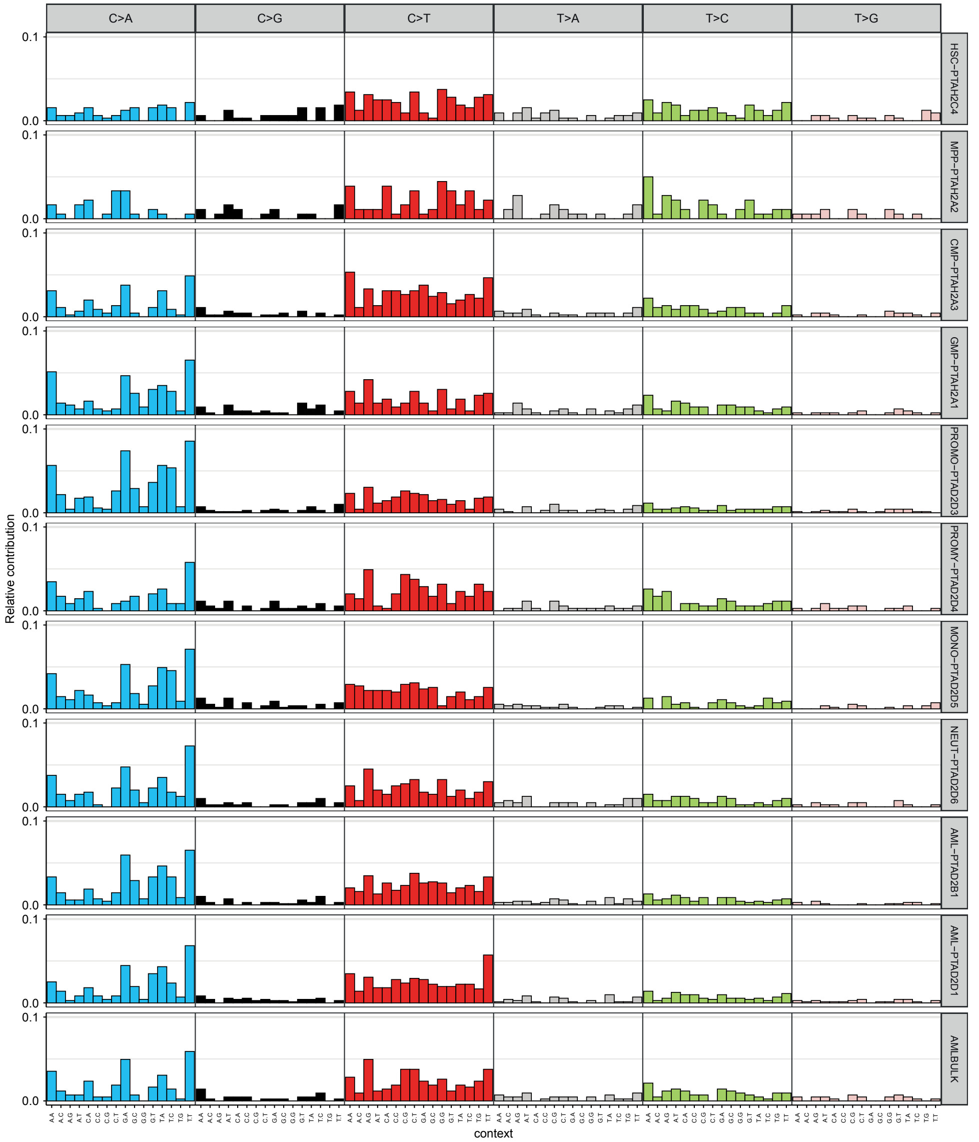


Figure S8: Differentiated AML cells and blasts have a similar mutation profile.
Relative contribution of each trinucleotide change to the point mutation spectra of differentiated AML cells and AML blasts (HSC-PTAH2C4: n = 392; MPP-PTAH2A2: n = 221; CMP-PTAH2A3: n = 555; GMP-PTAH2A1: n = 523; PROMO-PTAD2D3: n = 1033; PROMY-PTAD2D4: n = 432; MONO-PTAD2D5: n = 817; NEUT-PTAD2D6: n = 494; AML-PTAD2B1: n = 803; AML-PTAD2D1: n = 836; AMLBULK: n = 473).

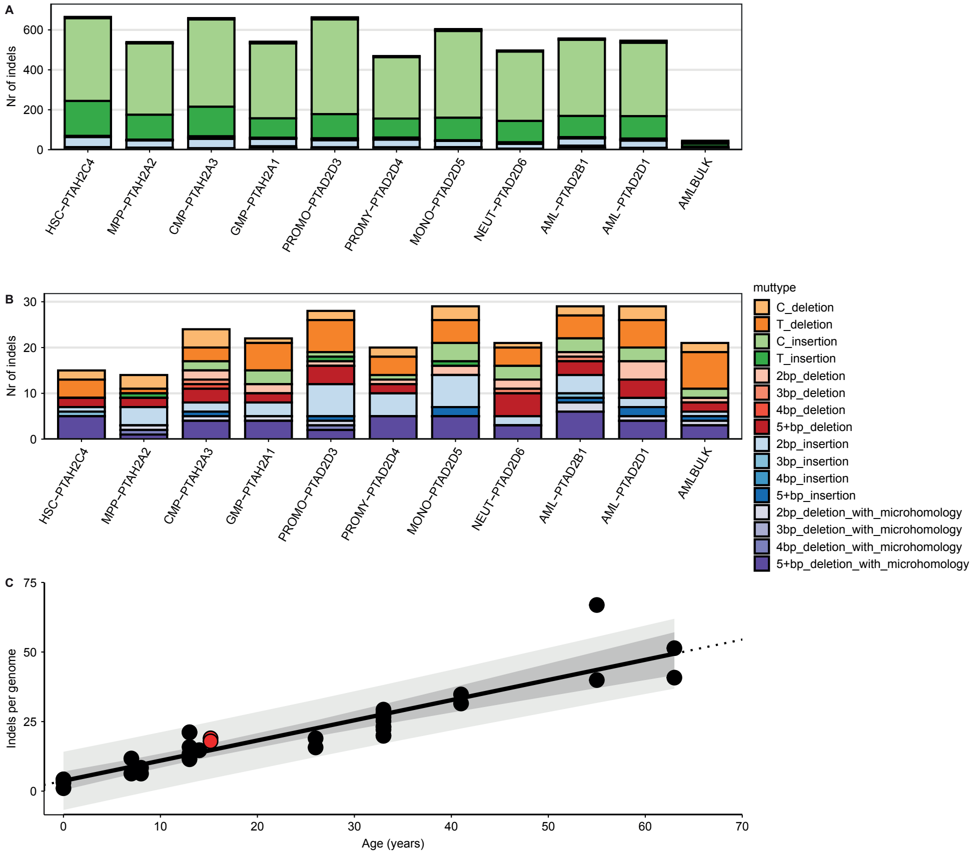


Figure S9: PTATO identified the mutation patterns and loads of indels in an AML patient.
a Absolute contribution of the indicated mutation types to the indel spectrum for different cell types in an AML patient before filtering with PTATO. **b** Absolute contribution of the indicated mutation types to the indel spectrum for different cell types in an AML patient after filtering with PTATO. **c** The number of indels is plotted against the age in years of the donor. Each dot is a sequenced sample. The black line shows the mean fitted number of indels at that age (linear mixed-effects model). The dark grey area shows the 95% confidence interval of the model, while the light grey area shows the 95% prediction interval. Black dots are clonally expanded HSCPs from normal donors (n = 33; donors = 11), whereas the two red dots are an HSC (top) and MPP (bottom) from the AML patient.

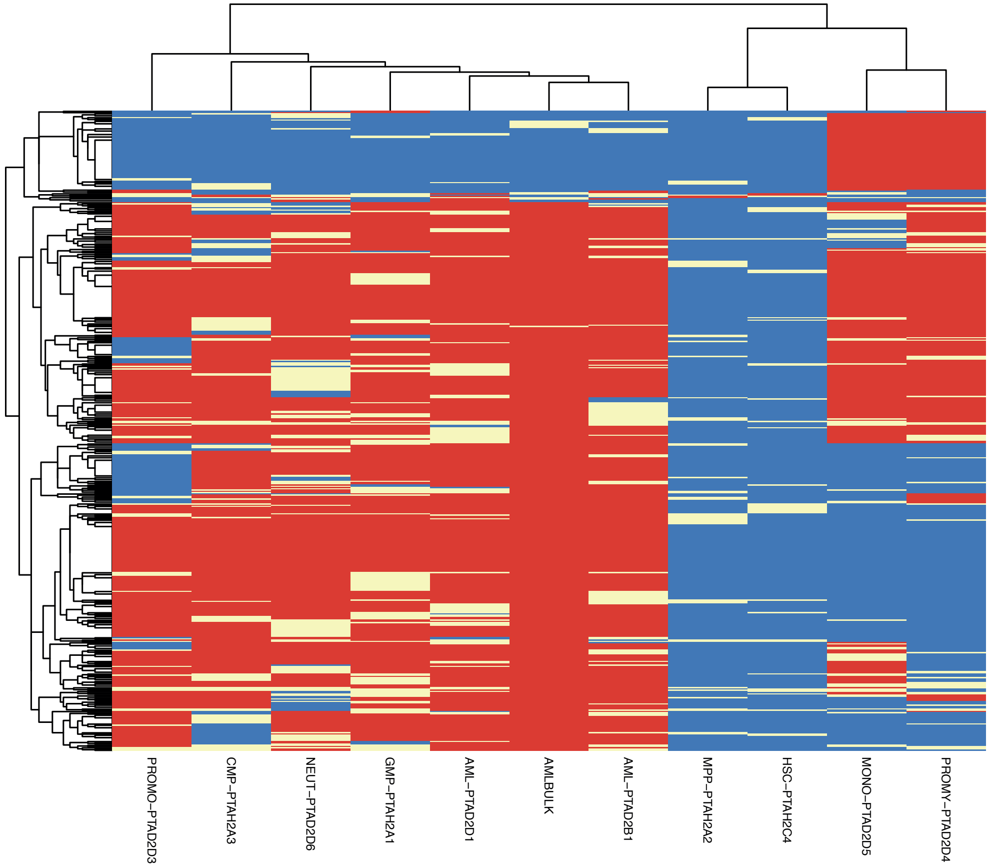


Figure S10: Many mutations are shared between differentiated AML cells and blasts.
Heatmap showing the somatic substitutions shared between multiple samples of an AML donor. Each row is a substitution and each column is a sample. Red indicates that a variant is present, blue indicates that a variant is absent and yellow indicates that the variant failed the quality control.

Table S1: Overview of the used samples

Subject	Sample	Label	Sample type	Type	Purpose	Sequencing	Sample source	Sequencing depth
IBFM35	IBFM35-DX2BM-AMLBULK	NA	Bulk	AML	Bulk control	Internal	GPOH	29
IBFM35	IBFM35-DX2BM-HSCPTAP1D9	IBFM35_1	PTA	HSC/Blast	Training	Internal	GPOH	17.4
IBFM35	IBFM35-DX2BM-HSCPTAP1E9	IBFM35_2	PTA	HSC/Blast	Training	Internal	GPOH	15.5
IBFM35	IBFM35-DX2BM-HSCPTAP1G9	IBFM35_3	PTA	HSC/Blast	Training	Internal	GPOH	16.3
IBFM35	IBFM35-DX2BM-MSCBULK	NA	Bulk	MSC	Bulk control	Internal	GPOH	33.4
IBFM26	IBFM26-DX2BM-AMLBULK	NA	Bulk	AML	Bulk control	Internal	GPOH	36.9
IBFM26	IBFM26-DX2BM-HSCPTAP1B8	IBFM26_1	PTA	HSC/Blast	Training	Internal	GPOH	18.3
IBFM26	IBFM26-DX2BM-TCELLBULK	NA	Bulk	T-cell	Bulk control	Internal	GPOH	37.4
Pt1	Pt1AMLBULK	AMLBULK	Bulk	AML	Bulk control	Internal	PMC	34.3
Pt1	Pt1MSCBULK	NA	Bulk	MSC	Bulk control	Internal	PMC	38.5
Pt1	Pt1-BMAML-PTAD2B1	AML-PTAD2B1	PTA	AML blast	Training	Internal	PMC	17
Pt1	Pt1-BMAML-PTAD2D1	AML-PTAD2D1	PTA	AML blast	Training	Internal	PMC	18.6
Pt1	Pt1-BMCMP-PTAH2A3	CMP-PTAH2A3	PTA	CMP	AML insight	Internal	PMC	17.9

Pt1	Pt1-BMGMP-PTAH2A1	GMP-PTAH2A1	PTA	GMP	AML insight	Internal	PMC	16.2
Pt1	Pt1-BMHSC-PTAH2C4	HSC-PTAH2C4	PTA	HSC	AML insight	Internal	PMC	16
Pt1	Pt1-BMMONO-PTAD2D5	MONO-PTAD2D5	PTA	Monocyte	AML insight	Internal	PMC	17.4
Pt1	Pt1-BMMPP-PTAH2A2	MPP-PTAH2A2	PTA	MPP	AML insight	Internal	PMC	16.2
Pt1	Pt1-BMNEUT-PTAD2D6	NEUT-PTAD2D6	PTA	Neutrophil	AML insight	Internal	PMC	16
Pt1	Pt1-BMPROMO-PTAD2D3	PROMO-PTAD2D3	PTA	Promonocyte	AML insight	Internal	PMC	19.1
Pt1	Pt1-BMPROMY-PTAD2D4	PROMY-PTAD2D4	PTA	Promyelocyte	AML insight	Internal	PMC	18
PMCCB15	PMCCB15-CBMPP-PTAP6F4	PMCCB15_1	PTA	Cord blood MPP	Training	Internal	PMC	20.2
PMCCB15	PMCCB15-CBMPP-PTAP6F5	PMCCB15_2	PTA	Cord blood MPP	Training	Internal	PMC	17.8
PMCCB15	PMCCB15-CBMPP-PTAP6F6	PMCCB15_3	PTA	Cord blood MPP	Training	Internal	PMC	20.2
PMCCB15	PMCCB15-CBWTVCR-HSP3L19	NA	Bulk	Cord blood HSC	Bulk control	Internal	PMC	15.2
PMCAHH1-FANCCKO	PMCAHH1-FANCCKO-C02B-03SC03E05-PTAP1D7	PMCAHH1-FANC-CKO_1	PTA	FANCC-knockout	Training	Internal	PMC	14
PMCAHH1-FANCCKO	PMCAHH1-FANCCKO-C02B-03SC03E05	NA	Bulk	FANCC-knockout	Subclone	Internal	PMC	16.7
PMCAHH1-FANCCKO	PMCAHH1-FANCCKO-C02B03	NA	Bulk	FANCC-knockout	Bulk control	Internal	PMC	17.9
ENU-1-1	Cord_blood_donor	ENU-1-1	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	18.9
ENU-1-2	Cord_blood_donor	ENU-1-2	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	19.5
ENU-1-3	Cord_blood_donor	ENU-1-3	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	24.3
ENU-1-4	Cord_blood_donor	ENU-1-4	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	17.2
ENU-1-5	Cord_blood_donor	ENU-1-5	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	20.4
ENU-2-1	Cord_blood_donor	ENU-2-1	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	17
ENU-2-2	Cord_blood_donor	ENU-2-2	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	19.2
ENU-2-3	Cord_blood_donor	ENU-2-3	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	19.8
ENU-2-4	Cord_blood_donor	ENU-2-4	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	14.1
ENU-2-5	Cord_blood_donor	ENU-2-5	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	17.9
ENU-3-1	Cord_blood_donor	ENU-3-1	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	20.7
ENU-3-2	Cord_blood_donor	ENU-3-2	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	18.2
ENU-3-3	Cord_blood_donor	ENU-3-3	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	15.5
ENU-3-4	Cord_blood_donor	ENU-3-4	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	16.6
MAN-2-1	Cord_blood_donor	MAN-2-1	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	17
MAN-2-2	Cord_blood_donor	MAN-2-2	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	14.7
MAN-2-3	Cord_blood_donor	MAN-2-3	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	18.9
MAN-2-4	Cord_blood_donor	MAN-2-4	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	15.2
MAN-2-5	Cord_blood_donor	MAN-2-5	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	22
MAN-3-1	Cord_blood_donor	MAN-3-1	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	19.6
MAN-3-2	Cord_blood_donor	MAN-3-2	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	17
MAN-3-3	Cord_blood_donor	MAN-3-3	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	22.1
MAN-3-4	Cord_blood_donor	MAN-3-4	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	19.8
MAN-3-5	Cord_blood_donor	MAN-3-5	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	22.9
VHC-0-1	Cord_blood_donor	VHC-0-1	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	15.7
VHC-0-2	Cord_blood_donor	VHC-0-2	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	23.9
VHC-0-3	Cord_blood_donor	VHC-0-3	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	19.8
VHC-0-4	Cord_blood_donor	VHC-0-4	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	19.3
VHC-0-5	Cord_blood_donor	VHC-0-5	PTA	Cord blood HSPC	Validation	Ref: 5	Ref: 5	21.8
BULK-1	Cord_blood_donor	BULK-1	Bulk	Cord blood HSPC	Bulk control	Ref: 5	Ref: 5	39.9

Chapter 6

General discussion

Freek Manders[1]

[1] Princess Máxima Center for Pediatric Oncology and Oncode Institute, Heidelberglaan 25,
3584CS Utrecht, The Netherlands

Introduction

Somatic mutations accumulate in the human body throughout life[1–16]. They play a role in several diseases such as neurogenerative disorders and diseases associated with aging, as discussed in **chapter 1**[17–20]. Additionally, they can drive cancers and other clonal expansion disorders such as clonal hematopoiesis of indeterminate potential, which can itself also develop further into cancer[21–25]. However, somatic mutations can be difficult to detect, because their generally low allele fraction in bulk tissues makes them difficult to distinguish from technical noise[26]. Furthermore, for many of these somatic mutations, the processes that cause them are unknown[27,28]. The goal of this thesis was to discover previously unidentifiable somatic mutations in normal cells and to improve methods for analyzing patterns within these mutations in order to better understand the processes that caused them. To do this, we first set out to measure somatic mutations in fetuses with and without Down syndrome. Some of these mutations occurred as early as the first cell division, complicating their detection. We also developed new methods to identify mutagenic mechanisms by in-depth mutational analyses and applied these on a human cell line in which we deleted various DNA repair genes. In addition, we studied the lifelong mutation accumulation in mitochondria and compared the mitochondrial genomes of normal and cancer cells. Finally, we developed a toolkit to distinguish somatic mutations from artifacts in single cells that were sequenced with PTA, allowing the genomes of differentiated cells to be analyzed[29]. Together, these projects improved our understanding of the patterns of mutations in normal cells and through this understanding brought us closer to understanding the origin of cancer.

The impact of the fetal mutation rate

Several studies have found that the mutations driving pediatric cancers like leukemia and rhabdoid tumors can already occur during fetal or possibly even embryonic development[30–36]. In the work described in **chapter 2**, we investigated the origin and abundance of these early somatic mutations in normal fetal stem and progenitor cells.

Normally, somatic mutations are identified by comparing a sample with a bulk tissue control[25,37]. However, somatic mutations can be present at observable levels in bulk tissues if they occur early enough in the development[38,39]. Therefore, instead of just using a bulk control, we sequenced multiple clonally expanded single cells and used them as a control for each other. As a result, we could observe somatic mutations that occurred before gastrulation and might have originated as early as the first cell division of the fertilized egg cell. In the future, this type of approach might be used to detect early occurring somatic mutations in pediatric cancers that are diagnosed early after birth and that are suspected of having an embryonic origin[31].

Having identified the somatic mutations in fetal stem and progenitor cells, we observed that hematopoietic stem and progenitor cells (HSPCs), which are difficult to distinguish by sorting and have similar mutation loads[10], have a 5.8 times higher somatic mutation rate during fetal development compared to the postnatal rate. This higher mutation rate might contribute to the increased incidence of leukemias in children compared to young adults[40]. However, we also observed an increased mutation rate in fetal intestinal stem cells, even though children do not have an increased risk of developing intestinal cancers. These observations support findings that other factors besides the mutation rate, such as epigenetic or developmental changes, might play a role in the development of pediatric cancers[31]. The importance of such factors is exemplified by the finding that in newborns only 1% of pre-leukemic clones with the ETV6-RUNX1 fusion develop into a full acute lymphoblastic leukemia, showing that the driver alone is not sufficient for the development of cancer[30]. The importance of the increased mutation rate of cells during development on cancer is thus not yet fully clear, as was discussed in more detail in **chapter 1**.

To obtain more insight into the mutational processes causing the observed mutations, we analyzed the mutational spectra and observed that the mutations in HSPCs could be explained by mutational signatures SBS1 and SBS5, both of which are clock-like signatures that are known to accumulate in a linear fashion with age[1,41]. We observed the same signatures in pediatric acute myeloid leukemia (AML) blasts, indicating that no additional mutational processes are necessary to initiate AML. However, many drivers causing pediatric leukemias are structural variants, which were not present in the fetal HSPCs[42,43].

Another way to study the link between somatic mutations in normal cells and carcinogenesis is by studying individuals at risk of developing cancer. Young children with Down-syndrome have a 500-fold increased risk of developing Down-syndrome associated acute megakaryoblastic leukemia (DS-AMKL) and a 7-20 times higher risk of developing acute lymphoblastic leukemia[44,45]. DS-AMKL is preceded by transient abnormal myelopoiesis, which is already present at birth and characterized by GATA1 driver mutations and thus likely originates during fetal development[34,45]. As described in **chapter 2**, we investigated stem and progenitor cells of fetuses with Down-syndrome and observed 34 extra substitutions compared to regular fetuses. These mutations were caused by the same mutational processes as the mutations in regular fetuses, which could contribute to their increased chance of developing leukemias. However, this increase in mutation load seems too small to fully explain the increased leukemia incidence in young children with Down-syndrome, suggesting that epigenetic and gene expression differences also play a role[45,46].

Mutation accumulation in normal stem cells

Investigating the mutations in normal cells throughout life allows us to better understand the accumulation of mutations and how this contributes to disease. By knowing what processes drive mutagenesis in normal cells, it can determined if and how they are perturbed in diseases, such as cancer. Sequencing normal clonally expanded single cells, such as the fetal HSPCs we analyzed, allows for a direct comparison with cancer, because cancers are also clonally expanded from an original transformed single cell. The mutation accumulation of the stem cells of many tissues have been assessed in the last few years, however, there are still several unresolved questions[1-16].

As described in **chapter 2**, we observed an increased mutation rate in stem and progenitor cells during fetal development, caused by an increased contribution of SBS1 and SBS5. It is currently not known why the mutational processes behind these two signatures are more active during fetal development, but several possibilities, like an increased cell division rate, were discussed in **chapter 1**. The increased mutation rate during fetal development contrasts with the mutation rate during adult life, which is remarkably constant in all the tissues that have been investigated so far. Most studies into somatic mutation rates during life have used donors with ages below 80. Older donors have been sequenced, but these studies either did not sequence single cells or they used single-cell multiple displacement amplification, which is not very accurate[47,48]. It is possible that the mutations that have accumulated during life in older donors might affect the mutation rate of their stem cells. On the other hand, it is also possible that centenarians have reached their old age, because they were protected from developing cancers by a below average mutation rate, caused by germline variants or other unknown protective factors. Studies into the blood or other tissues of centenarians and super centenarians might elucidate if the mutation rate of stem cells is changed at a very old age, just like it is at a very young age.

While it is now known that somatic mutations accumulate linearly with age in stem cells, it is not yet known what causes the variation between patients. Understanding this variation could help us better understand the mutational processes generating the mutations in stem cells. Some of the variation might be caused by germline variants[49], however this cannot explain all the variance as we and others have observed that there are even differences in mutation load between stem cells of a single tissue of a single donor[5,10,50]. The variance in mutation load is thus likely caused at least in part by differences between cells at an intra-individual level. These environmental factors could be investigated in the future by methods combining single-cell genomics and transcriptomics to see if cells with a higher mutation load have a lower expression of DNA-repair genes, if they have up- or down-regulated certain pathways that influence the mutation rate[51]. Additionally, spatial genomics methods could be used to see how the physical location of a cell influences its mutation rate. Cells located closer to the border of a tissue or micro-environment might, for example, have a higher exposure to oxidative stress or other mutagenic factors.

The lung stem cells of smokers have an increased mutation load compared to non-smokers; however, in a subset of cells of ex-smokers this increase has disappeared[3]. Since mutations cannot be repaired anymore after being fixed in the genome, this indicates that the affected cells are being replaced by other stem cells. One possible explanation for this phenomenon is the concept of quiescent or long-term stem cells[52,53]. This hypothesis states that most stem cells in some tissues are only alive for several months to years and that they are replenished by longer lived long-term stem cells when necessary. These long-term stem cells, which can be quiescent for prolonged periods, would then also be more protected from damage to their genome. Since these long-term stem cells are less numerous and more protected than the short-term stem cells, they have likely been rarely sequenced by studies into the mutation accumulation of normal cells. Identifying these long-term stem cells, if they exist, and investigating their mutational loads might help us better understand mutation accumulation in normal cells.

Mutation accumulation in differentiated cells

While the mutation accumulation of adult stem cells of different human tissues has been analyzed[1–16], most differentiated cell types have not yet been assessed due to technical limitations. In the coming years novel single cell sequencing technologies, that do not depend on the clonal expansion of cells will be used to analyze the mutations in these single cells[29]. Knowing the mutation load of more differentiated cell types will help us understand cancer, as we can compare the mutation load of differentiated cells to cancers that developed, via a process of dedifferentiation, from a differentiated cell type[24,54–56]. An example of this is the finding that melanocytes can dedifferentiate during melanoma tumorigenesis[54]. Additionally, sequencing differentiated cells can also aid in our understanding of AML, as it has been suggested that differentiation is necessary for the development of this type of cancer[57]. Knowing the mutation load of differentiated cells can also help us better understand other diseases. The neurons of Alzheimer's patients, for example, were recently shown to have an increased mutation load compared to normal neurons, which might impair their function and contribute to the development of the disease[58].

Differentiated cells have a limited lifespan and division potential, because of their low telomerase activity[59,60]. As a result, they are expected to be less likely to transform into cancer cells and any damage to their genomes that affects their functioning will only be inherited by a limited number of transient progeny. Because of this, it has been hypothesized that the genomes of differentiated cells need to be less protected than that of stem cells and that therefore, differentiated cells could have a higher mutation rate than stem cells, caused by either a decreased activity of the DNA repair machinery or a less protective niche[59,61,62]. In the coming years, single-cell and/or molecule genome sequencing will likely be used to find out if this hypothesis is valid. Some studies have already observed differences in mutation loads between differentiated cells and stem cells, but few systematic analyses using high-quality data have been performed so far[63–65]. Even with single-cell whole genome sequencing, identifying the mutation rates of differentiated cells might be difficult, because these cells often have a limited lifespan and therefore most of the mutations in them likely occurred before their differentiation. As a result, it might be necessary to sequence cells of which the time since differentiation is relatively long and either known or approximated.

Single-cell whole genome sequencing techniques will also be used to investigate if the process of differentiation itself might cause a burst of mutations. The large transcriptomic and epigenetic changes that constitute differentiation might damage the genome through oxidative or other stresses[66]. One known example of mutations caused by the differentiation process itself, is the somatic hypermutation caused by the activation-induced cytidine deaminase associated with the V(D)J-recombination in B-cells and T-cells[48,63,67]. However, whether the process of differentiation also causes mutations in other cell types is not yet clear.

Next to the mutation rates, the mutational processes causing mutations can also be investigated in differentiated cells. It is possible that compared to stem cells, differentiated cells contain dissimilar contributions of some mutational processes or even contributions of distinct mutational processes, which are not present in the matching stem cells. Changes in the mutational processes of cells after differentiation could be caused by a lack of DNA-repair machinery, the differentiation process itself, or an increased exposure to environmental factors because of a less protective niche[59,61,62].

Overall, while our understanding of mutation accumulation in normal cells has improved a lot the last few years, there are still many unanswered questions. Aided by the continuous development of novel technologies, we will likely begin answering these questions in the next few years.

Improving the identification of mutations in single cells

Analyzing the genomes of differentiated cells requires single-cell sequencing technologies that do not depend on a cells' self-renewal ability. Historically, these technologies had a low sensitivity and accuracy[29]. However, several methods have recently been developed like single-molecule duplex sequencing (NanoSeq), multiplexed end-tagging amplification of complementary strands (META-CS), and primary template-directed amplification (PTA) that claim to have a significantly improved sensitivity and precision[29,64,68]. For the project described in **chapter 5**, we focused on PTA, because it has a high sensitivity, is relatively easy to perform, and is supported by a commercial kit. Additionally, it distinguishes between individual cells, which NanoSeq does not do. PTA results in less artifacts and has a much more even genome coverage compared to its predecessor multiple displacement amplification[29].

In order to analyze PTA data and accurately identify somatic mutations, we developed the PTA analysis toolkit (PTATO). We showed that PTATO could identify single base substitutions (SBSs), insertions and deletions (indels), and structural variants (SVs) with high sensitivity and precision. Additionally, we showed that PTATO could accurately estimate the mutation load of cells and analyze the mutational processes that had been active in them. Finally, we applied PTATO on samples from an AML patient to illustrate how it could be used to gain new biological insights.

While PTATO filters out most artifacts, we still have about a 20% false positive rate and a 20% false negative rate for the detection of SBSs. This error rate could result in cancer drivers being missed. Additionally, one cannot be confident that any drivers that were found in only one cell are real. There are several potential approaches to improve the performance of PTATO. First, PTATO was trained on a relatively small dataset of ten samples. Training PTATO on a larger dataset, with samples from multiple cell types and cell lines that have been exposed to a variety of different mutational processes could increase its accuracy and stability. Combining data from multiple labs, as it becomes available, to create a single large training set could further increase PTATO's stability. A larger dataset would also allow us to replace the random forest with a more powerful and flexible model like a neural network. A convolutional network in particular might work well on the features describing the ten bases upstream and downstream of a potential mutation[69,70]. The outcome of the convolutional layers could then be combined with the other features in regular full layers to give the final prediction.

A larger dataset for training would also make it possible to use more features without overfitting. These could be epigenetic features like Ensembl's regulatory regions or the number of bases upstream and downstream of potential mutations could be increased[71,72].

While these methods could improve PTATO, there might be an upper limit to its accuracy. It might not be feasible to distinguish artifacts from real mutations that are very similar to artifacts. PTA, for example, generates a lot of artifactual C insertions at C repeat regions, making it difficult to detect real C insertions at these locations. Furthermore, any mutations that occur during the early steps of the PTA protocol might be indistinguishable from real mutations if they occur outside of typical artifact locations.

Currently, work is underway to combine PTA with single-cell RNA sequencing[51]. Since this combined method still requires variant filtration, PTATO can probably be used for this, in combination with a single-cell transcriptomics pipeline. Integrating single-cell genomics and transcriptomics will help clarify the differences in mutation accumulation between different types of cells and will help elucidate the effects that both coding and non-coding mutations have on gene expression. Furthermore, this combined method could be used to investigate the mutation rate of differentiated cells, since it is possible to use the transcriptomic profile of a cell to approximate how far along its differentiation trajectory it is.

One major downside of both PTA and whole genome sequencing in general is the relatively low number of cells that can be analyzed, as each cell requires a separate sample. While sequencing costs have gone down over time, sequencing the genomes of thousands of cells at a time, similar to single-cell RNA sequencing experiments, is unlikely to become affordable soon[73,74].

Mutagenesis in mitochondria

Next to the nuclear genome, mutations can also occur in the mitochondrial genome. In the investigation described in **chapter 4**, we characterized mutation accumulation in mitochondrial DNA (mtDNA). We observed that mtDNA accumulates mutations with age in normal cells and by comparing our data with cancer samples we found that the majority of mutations in cancer were the result of pre-malignant mutation accumulation[75]. However, this was tissue type dependent as colon cancers showed a larger increase in the mutation load compared to blood cancers. It is not clear what caused this difference, but one explanation could be a higher activitiy of mitochondria in colon cells, as indicated by the higher mtDNA copy number load. It is also not yet known how big the difference in mitochondrial mutation load is between normal cells and cancers in other tissue types. Analyzing the mitochondrial genomes of more tissue types might provide hints as to what causes the increased mutation load in cancer. Focusing on specific samples that are known to harbor dysfunctional mitochondria might also be fruitful. Next to the small increase in mutation load in cancer, we also found that different types of cancer treatment did not result in an increased mutation load, suggesting that these treatments do not cause a risk to mitochondrial genomes[76,77]. One potential explanation for this is the ability of cells to remove damaged mitochondrial genomes.

Mitochondrial genomes have been extensively linked to cancer and other diseases[75,76,78–82]; however, our results would suggest that mitochondrial mutations do not play a large role in the origin of cancer. One explanation for this is that mitochondria can play a role in cancer without containing DNA damage. Many mitochondrial proteins are encoded in the nucleus and mutations in them could lead to incorrectly functioning mitochondria without the mitochondrial DNA itself being damaged[76]. Additionally, an aberrant expression of genes regulating the mitochondria could change a cells metabolism. This also matches the fact that a single cell contains many mitochondrial genomes[83,84]. A mutation in only one of these genomes is unlikely to significantly affect a cells metabolism unless its VAF was first increased through either selection or random drift[78].

It is also possible that we are missing a considerable number of mitochondrial mutations. We sequenced clonally expanded single cells. During this clonal expansion the variant allele frequency of mutations could decrease because of selection or neutral drift, causing them to not be detected. However, we do not expect that the limited time of the clonal expansions will massively impact the VAF of most mutations. Additionally, while mutations with a VAF below the detection limit will be missed, they are unlikely to have a biological effect as mentioned previously. One way to better detect mitochondrial variants would be to directly sequence single cells without a clonal expansion, via a technique like PTA[29]. For the PTA analysis toolkit we described in **chapter 5**, we focused on the nuclear genome, but in the future, it would be interesting to combine it with the mitochondrial analyses described in **chapter 4**.

We analyzed mitochondrial genomes using whole genome sequencing data. It is also possible to sequence samples that have been enriched for mtDNA[85–87]. This is cheaper, because the nuclear genome is not sequenced; however, this also means a lot of genomic information is lost. The main benefit of the method we used is that it can be used on the large amount of whole genome sequencing data that has already been generated.

Overall, the importance of mitochondrial mutations is not clear yet. However, their importance might be elucidated by analyzing more samples and by directly sequencing single cells. Since the mitochondrial genome analysis pipeline we used only uses a small amount of computational resources, adding it to existing sequencing pipelines would not be very costly.

The processes behind somatic mutations

By observing patterns in the somatic mutations present in cells, it is possible to analyze the processes that caused them. In **chapter 3** we described the second version of the MutationalPatterns package to analyze these patterns. We included new mutation types like indels and double base substitutions. We also included functions to perform stricter and bootstrapped signature refitting, and functions to analyze lesion segregation and illustrated their use on cell lines in which we deleted specific DNA repair genes.

Recently, COSMIC published a set of mutational signatures for copy numbers[88]. Attempts to define mutational signatures for structural variants have also been made[89–91]. It would be good to extend MutationalPatterns to be able to analyze and plot copy number mutation patterns, since this would likely require a limited amount of work and might make MutationalPatterns an even more comprehensive tool to analyze the mutation accumulation of cells. Extending MutationalPatterns to structural variants would, however, be more difficult since there is less consensus on how they should be identified, filtered, and represented in VCF files[89,92]. Furthermore, no widely used set of canonical structural variant signatures currently exists.

The last few years has seen a large increase in the defined number of SBS mutational signatures, resulting in overfitting where signatures are incorrectly attributed to a sample[27,41,93–97]. We tried to tackle this issue with the stricter and bootstrapped refitting functions, however this is only moderately effective when fitting against the latest large signature sets, like the current set of COSMIC signatures (v3.3)[27]. Another approach that we have used is to perform NMF on a combination of the data of interest as well as a standard in-house dataset followed by signature refitting using the smaller set of signatures defined by the NMF. While this decreases false positives and gives more stable results, it does provide a bias for the signatures present in the standard in-house dataset[96]. Another method that has been proposed is to use tissue-specific signatures[96]. Finally, it is also possible to perform signature refitting against a small set of common signatures, followed by a second refit using rare signatures for samples that cannot be properly explained by the common signatures alone[94]. While these approaches might reduce the number of false positives, they also introduce bias[94,96]. The best method or combination of methods is not yet clear and might depend on the research question. The accuracy of signature refitting also depends strongly on the signatures to which a sample was exposed. Refitting performs well for signatures with large contributions from only a few features but performs somewhat poorly for more ‘flat’ signatures that contain small contributions from many features[50,96]. Overall, the accuracy of signature refitting is not ideal, and care must be taken when interpreting its results.

Mutational signatures were originally designed on a set of 21 breast cancers using the 96-trinucleotide context as its features[95]. Since then, signatures have been defined on ever larger datasets, but the features that are used for SBSs have remained the same[27,94]. With these larger datasets it is possible to use a large sequence context or to include other features like the distance to the nearest gene body, since we and others have shown that these features can influence mutational processes[27,98–100]. This could help distinguish mutational processes, that are very similar based on the current 96-trinucleotide context. Therefore, we allowed for a larger sequence context to be used in the second version of MutationalPatterns.

Another way to improve the features used to define signatures would be to make them more informative. Currently, the 96 features are all treated equally, while some are much more informative than others. When using a larger mutational context, this will become an even bigger issue, as adding a single base to both the the 5’ and 3’ ends, increases the number of features 16-fold. One potential way to solve this, would be to use an autoencoder or PCA to reduce the number of features to a smaller more informative amount[101]. This reduced number of features could be used as a type of signatures themselves or they could be further analyzed with NMF. For the latter option, a single autoencoder could be trained on a large pan-cancer dataset and the features it created could then be used by other labs to perform signature refitting or NMF. Overall, it might be good to define signatures using an extended feature set that has been treated with a dimensionality reduction method.

The patterns present in somatic mutations have been used for more than just analyzing the processes that caused them. In the work described in **chapter 5** we used mutation patterns to distinguish true mutations from artifacts. Additionally, mutation patterns have been used to estimate the cell-of-origin of cancers and to distinguish healthy people from cancer patients using the cell-free DNA present in plasma[102–105]. On top of this, more use-cases for mutational patterns might be developed in the future.

Impact of mutational processes

As described in **chapter 2**, we observed that SBS1 and SBS5 were present in normal fetal stem and progenitor cells and that no additional mutational signatures were needed for the development of AML. However, the importance of these two signatures to the development of AML is still unknown. This reflects a wider issue in the field, as it is difficult to determine how often different mutational processes result in mutations that drive cancers or affect a person's health. In the work described in **chapter 3**, we started to address this by adding a function to MutationalPatterns that predicts how likely a mutational process is to cause, missense and stop-gain mutations. However, while this provides a useful indication on how damaging a mutational process is, it doesn't show how often a process results in cancer drivers. To properly determine the impacts of mutational processes it is necessary to determine for individual mutations by which mutational process they were generated.

In the investigation described in **chapter 5**, we did something very similar to this, as we trained a random forest on an extended mutation context to predict if a mutation was a PTA artifact. A similar approach has been developed in our group to predict if individual mutations were caused by ganciclovir[106]. Extending these methods to other mutational signatures would allow individual driver mutations to be linked to mutational processes. By then applying these methods on large pan-cancer datasets the overall impact of mutational processes on cancer could be clarified. Next to pan-cancer analyses, these methods should also be applied per cancer type, because differences in drivers between cancer types will likely result in differences in the damage potential of mutational processes between tissues[25]. However, while this approach will work well on signatures that mutate nucleotides within very specific genomic contexts, it will work less well on 'flat' signatures that cause mutations within a less specific context.

The random forest we trained only had a specificity of around 75%, however this is not a major issue for identifying the impact of mutational processes. The probability that a driver mutation belongs to a specific mutational signature can be used as a weight when adding up the drivers linked to a signature, thus averaging out the uncertainties in these models.

Identifying the impact of different mutational processes on cancer will lead to a better understanding of the origin of cancer and could even be used for prevention. Knowledge on the impact of mutagenic chemotherapy treatments, for example, could aid clinicians in determining what treatments and therapy doses to use.

Concluding remarks

In this thesis I have focused on the genome wide identification of somatic mutations. While my focus was on base substitutions and indels in the nuclear genome, I also looked at structural variants in the works described in **chapter 2** and **chapter 5** and looked at mitochondrial variants in the investigation described in **chapter 4**. I could compare the mutation loads between samples and find patterns in the identified mutations, informing us about the processes that caused them. For the mutations in the coding part of the genome, I also determined whether they were likely to drive cancers.

However, since the vast majority of mutations that we identified are located outside of the coding region of the genome, their biological effects cannot be easily determined[25,107–109]. This is not just the case for somatic mutations, but also for germline variants[110]. While these variants do not code for proteins, they have been linked to many diseases including cancer by GWAS, eQTL, and other techniques[107–109,111–113]. Epigenomic studies have also shown that many non-coding regions of the genome are functional as enhancers, promoters, or other functional elements[71,114,115]. Multiple tools have been made to predict whether variants in non-coding genomes cause disease by, for example, using epigenetic data and by looking at how conserved different regions of the genome are[70,108,111,116]. However, so far these tools have had a limited accuracy[109]. While many variants are known to increase or decrease an individual's risk of a specific disease, there are still very few non-coding variants that can be identified as directly driving or causing a disease. An exception to this are mutations in the *TERT2* promoter, which are known to occur in over 50 cancer types[25,60]. Because of the difficulty in identifying the effects of non-coding mutations, cancer drivers are generally called within the exome[107].

Improved tools to determine the effects of all somatic mutations, and not just the ones in coding regions, would further increase the value of identifying somatic mutations in the whole genome, a major focus of this thesis, and would allow for new types of analyses on identified sets of somatic mutations. In my opinion, a better understanding of the impact on disease of somatic mutations and germline variants in the non-coding part of the genome is currently the most important challenge within the field of genetics.

In the future it might be possible to reduce mutation accumulation and with it the impact it causes on a person's health. Mutations caused by smoking, alcohol consumption, and eating red meat can be reduced with lifestyle changes, and the bans of mutagenic materials like chrome-6 in paint and asbestos have already prevented mutations and the cancers they would have caused[27,117–120]. Additionally, the ban of chlorofluorocarbons has reduced the damage to the earth's ozone layer, which has likely already lead to reduced mutation rates in the skin[121]. Furthermore, a future reduction in air pollution via stricter regulation and technological improvements could also prevent somatic mutations in the lungs from occurring.

A better understanding of mutational signatures commonly present in normal cells, like SBS1 and SBS5 could possibly also help to reduce their activity. Several germline variants have already been correlated with the activity of specific mutational signatures and more are likely to follow[49]. Furthermore, SBS5 was recently linked to REV1, so variants within this gene might be correlated to SBS5 exposure[122]. Gene editing in zygotes of variants associated with an increased mutation load could be used to reduce the rate with which they accumulate mutations during life. However, this is unlikely to happen in the next several decades. Overall, a future reduction in somatic mutation rates might reduce the prevalence of cancers and other diseases caused by somatic mutations.

In the coming years our knowledge of the mutation accumulation present in normal cells and the processes behind them will likely be further expanded. This will help us explain the origin of cancer and hopefully also provide us with hints on how to prevent it. Overall, the research described in this thesis has improved our understanding of mutation accumulation in normal cells and has produced tools that will hopefully be used to further improve our understanding in the coming years and thereby help elucidate the origin of cancer.

Acknowledgements

I would like to thank Ruben van Boxtel for providing me feedback.

References

1. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538(7624):260–4.
2. Franco I, Johansson A, Olsson K, Vrtacnik P, Lundin P, Helgadottir HT, et al. Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. Nat Commun. 2018;9(1):800.
3. Yoshida K, Gowers KHC, Lee-Six H, Chandrasekharan DP, Coorens T, Maughan EF, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature. 2020;578(7794):266–72.
4. Grossmann S, Hooks Y, Wilson L, Moore L, O'Neill L, Martincorena I, et al. Development, maturation, and maintenance of human prostate inferred from somatic mutations. Cell Stem Cell. 2021;1–13.
5. Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, et al. The mutational landscape of human somatic and germline cells. Nature. 2021;597(7876):381–6.
6. Park S, Mali NM, Kim R, Choi J-W, Lee J, Lim J, et al. Clonal dynamics in early human embryogenesis inferred from somatic mutation. Nature. 2021;597(7876):393–7.
7. Mitchell E, Spencer Chapman M, Williams N, Dawson KJ, Mende N, Calderbank EF, et al. Clonal dynamics of haematopoiesis across the human lifespan. Nature. 2022;606(7913):343–50.
8. Manders F, van Boxtel R, Middelkamp S. The Dynamics of Somatic Mutagenesis During Life in Humans. Vol. 2, Frontiers in Aging. 2021.
9. Lee-six H, Øbro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature. 2018;
10. Osorio FG, Rosendahl Huber A, Oka R, Verheul M, Patel SH, Hasaart K, et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. 2018 Nov;25(9):2308-2316.e4.
11. Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature. 2019;574(7779):538–42.
12. Franco I, Helgadottir HT, Moggio A, Larsson M, Vrta P, Johansson A, et al. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. 2019;1–22.
13. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;574(7779):532–7.
14. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019;565(7739):312–7.
15. J. LAR, Federico A, H. CTH, Yvette H, Laura O, Calli L, et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. Science (80-). 2020 Oct 2;370(6512):75–82.
16. Moore L, Leongamornlert D, Coorens THH, Sanders MA, Ellis P, Dentro SC, et al. The mutational landscape of normal human endometrial epithelium. Nature. 2020;580(7805):640–6.
17. Vijg J, Dong X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. Cell. 2020;182(1):12–23.
18. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. Nat Rev Genet. 2013;14(5):307–20.
19. Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. Somatic mosaicism: implications for disease and transmission genetics. Trends Genet. 2015;31(7):382–92.
20. Tiwari V, Wilson 3rd DM. DNA Damage and Associated DNA Repair Defects in Disease and Premature Aging. Am J Hum Genet. 2019;105(2):237–57.
21. Iñigo M, J. CP. Somatic mutation in cancer and normal cells. Science (80-). 2015 Sep 25;349(6255):1483–9.
22. Siddhartha J, L. EB. Clonal hematopoiesis in human aging and disease. Science (80-). 2019 Nov 1;366(6465):eaan4673.
23. Mustjoki S, Young NS. Somatic Mutations in "Benign" Disease. N Engl J Med. 2021;384(21):2039–52.
24. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discov. 2022 Jan 12;12(1):31–46.

25. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93.
26. Dou Y, Gold HD, Luquette LJ, Park PJ. Detecting Somatic Mutations in Normal Cells. Trends Genet. 2018;34(7):545–57.
27. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
28. Koh G, Degasperi A, Zou X, Momen S, Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer. 2021;21(10):619–37.
29. Veronica G-P, Sivaraman N, Yuntao X, David K, Robert C, Yakun P, et al. Accurate genomic variant detection in single cells with primary template-directed amplification. Proc Natl Acad Sci. 2021 Jun 15;118(24):e2024176118.
30. Greaves M. A causal mechanism for childhood acute lymphoblastic leukaemia. Nat Rev Cancer. 2018;18(8):471–84.
31. Filbin M, Monje M. Developmental origins and emerging therapeutic opportunities for childhood cancer. Nat Med. 2019;25(3):367–76.
32. Williams N, Lee J, Moore L, Joanna Baxter E, Hewinson J, Dawson KJ, et al. Phylogenetic reconstruction of myeloproliferative neoplasm reveals very early origins and lifelong evolution. bioRxiv. 2020;
33. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. Nature. 2020;578(7793):122–8.
34. Taub JW, Mundschau G, Ge Y, Poulik JM, Qureshi F, Jensen T, et al. Prenatal origin of GATA1 mutations may be an initiating step in the development of megakaryocytic leukemia in Down syndrome. Blood. 2004 Sep 1;104(5):1588–9.
35. Vitte J, Gao F, Coppola G, Judkins AR, Giovannini M. Timing of Smarcb1 and Nf2 inactivation determines schwannoma versus rhabdoid tumor development. Nat Commun. 2017;8(1):300.
36. Marshall GM, Carter DR, Cheung BB, Liu T, Mateos MK, Meyerowitz JG, et al. The prenatal origins of cancer. Nat Rev Cancer. 2014;14(4):277–89.
37. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun. 2015;6(1):10001.
38. Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature. 2014;513(7518):422–5.
39. Ju YS, Martincorena I, Gerstung M, Petljak M, Alexandrov LB, Rahbari R, et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature. 2017;543(7647):714–8.
40. de Magalhães JP. How ageing processes influence cancer. Nat Rev Cancer. 2013;13(5):357–65.
41. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio S a JR, Behjati S, Biankin A V, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.
42. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018 Feb;
43. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature. 2018;555(7696):321–7.
44. Hasle H, Clemmensen IH, Mikkelsen M. Risks of leukaemia and solid tumours in individuals with Down's syndrome. Lancet. 2000 Jan;355(9199):165–9.
45. Khan I, Malinge S, Crispino J. Myeloid leukemia in Down syndrome. Crit Rev Oncog. 2011;16(1–2):25–36.
46. Labuhn M, Perkins K, Matzk S, Varghese L, Garnett C, Papaemmanuil E, et al. Mechanisms of Progression of Myeloid Preleukemia to Transformed Myeloid Leukemia in Children with Down Syndrome. Cancer Cell. 2019 Aug;36(2):123-138.e10.
47. Holstege H, Pfeiffer W, Sie D, Hulsman M, Nicholas TJ, Lee CC, et al. Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. Genome Res. 2014 May 1;24(5):733–42.
48. Zhang L, Dong X, Lee M, Maslov AY, Wang T, Vijg J. Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. 2019;2–7.
49. Liu Y, Gusev A, Heng YJ, Alexandrov LB, Kraft P. Somatic mutational profiles and germline polygenic risk scores in human cancer. Genome Med. 2022;14(1):14.
50. Hasaart KAL, Manders F, van der Hoorn M-L, Verheul M, Poplonski T, Kuijk E, et al. Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis. Sci Rep. 2020;10(1):12991.
51. Zawistowski JS, Salas-González I, Morozova T V, Blackinton JG, Tate T, Arvapalli D, et al. Unifying genomics and transcriptomics in single cells with ResolveOME amplification chemistry to illuminate oncogenic and drug resistance mechanisms. bioRxiv. 2022 Jan 1;2022.04.29.489440.
52. Brunet de la Grange P, Vlaski M, Duchez P, Chevaleyre J, Lapostolle V, Boiron J-M, et al. Long-term repopulating hematopoietic stem cells and "side population" in human steady state peripheral blood. Stem Cell Res. 2013;11(1):625–33.
53. van Velthoven CTJ, Rando TA. Stem Cell Quiescence: Dynamism, Restraint, and Cellular Idling. Cell Stem Cell. 2019;24(2):213–25.
54. Köhler C, Nittner D, Rambow F, Radaelli E, Stanchi F, Vandamme N, et al. Mouse Cutaneous Melanoma Induced by Mutant BRaf Arises from Expansion and Dedifferentiation of Mature Pigmented Melanocytes. Cell Stem Cell. 2017;21(5):679-693.e6.
55. Perekatt AO, Shah PP, Cheung S, Jariwala N, Wu A, Gandhi V, et al. SMAD4 Suppresses WNT-Driven Dedifferentiation and Oncogenesis in the Differentiated Gut Epithelium. Cancer Res. 2018 Sep 4;78(17):4878–90.
56. Shih I-M, Wang T-L, Traverso G, Romans K, Hamilton SR, Ben-Sasson S, et al. Top-down morphogenesis of colorectal tumors. Proc Natl Acad Sci. 2001 Feb 27;98(5):2640–5.
57. Ye M, Zhang H, Yang H, Koche R, Staber PB, Cusan M, et al. Hematopoietic Differentiation Is Required for Initiation of Acute Myeloid Leukemia. Cell Stem Cell. 2015 Nov 5;17(5):611–23.
58. Miller MB, Huang AY, Kim J, Zhou Z, Kirkham SL, Maury EA, et al. Somatic genomic changes in single Alzheimer's disease neurons. Nature. 2022;604(7907):714–22.
59. Wyles SP, Brandt EB, Nelson TJ. Stem cells: the pursuit of genomic stability. Int J Mol Sci. 2014;15(11):20948–67.
60. Bell RJA, Rube HT, Xavier-Magalhães A, Costa BM, Mancini A, Song JS, et al. Understanding TERT Promoter Mutations: A Common Path to Immortality. Mol Cancer Res. 2016 Apr 14;14(4):315–23.
61. Choi E-H, Yoon S, Koh YE, Seo Y-J, Kim KP. Maintenance of genome integrity and active homologous recombination in embryonic stem cells. Exp Mol Med. 2020;52(8):1220–9.
62. Eliasson P, Jönsson J-I. The hematopoietic stem cell niche: Low in oxygen but a nice place to be. J Cell Physiol. 2010 Jan 1;222(1):17–22.
63. Machado H, Mitchell E, Obro N, Kubler K, Davies M, Maura F, et al. Genome-wide mutational signatures of immunological diversification in normal lymphocytes. bioRxiv. 2021;
64. Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing S V, Ellis P, et al. Somatic mutation landscapes at single-molecule resolution. Nature. 2021;593(7859):405–10.
65. Brandsma AM, Bertrums EJM, van Roosmalen MJ, Hofman DA, Oka R, Verheul M, et al. Mutation Signatures of Pediatric Acute Myeloid Leukemia and Normal Blood Progenitors Associated with Differential Patient Outcomes. Blood Cancer Discov. 2021 Sep 1;2(5):484 LP – 499.
66. Shen X, Wang R, Kim MJ, Hu Q, Hsu C-C, Yao J, et al. A Surge of DNA Damage Links Transcriptional Reprogramming and Hematopoietic Deficit in Fanconi Anemia. Mol Cell. 2020 Dec 17;80(6):1013-1024.e6.
67. Kasar S, Kim J, Improgo R, Tiao G, Polak P, Haradhvala N, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat Commun. 2015;6(1):8866.
68. Dong X, Longzhi T, Chi-Han C, Heng L, Sunney XX. Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. Proc Natl Acad Sci. 2021 Feb 23;118(8):e2013106118.
69. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and

RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8.
70. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 2015;12(10):931–4.
71. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. Genome Biol. 2015 Mar;16:56.
72. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020;48:D682–8.
73. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genet Med. 2018;20(10):1122–30.
74. Schwarze K, Buchanan J, Fermont JM, Dreau H, Tilley MW, Taylor JM, et al. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. Genet Med. 2020;22(1):85–94.
75. Yuan Y, Ju YS, Kim Y, Li J, Wang Y, Yoon CJ, et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nat Genet. 2020;52(3):342–52.
76. Zong W-X, Rabinowitz JD, White E. Mitochondria and Cancer. Mol Cell. 2016 Mar 3;61(5):667–76.
77. Ma K, Chen G, Li W, Kepp O, Zhu Y, Chen Q. Mitophagy, Mitochondrial Homeostasis, and Cell Fate. Vol. 8, Frontiers in Cell and Developmental Biology. 2020.
78. Schon EA, DiMauro S, Hirano M. Human mitochondrial DNA: roles of inherited and somatic mutations. Nat Rev Genet. 2012 Dec;13(12):878–90.
79. Alston CL, Rocha MC, Lax NZ, Turnbull DM, Taylor RW. The genetics and pathology of mitochondrial disease. J Pathol. 2016/11/02. 2017 Jan;241(2):236–50.
80. Greaves LC, Reeve AK, Taylor RW, Turnbull DM. Mitochondrial DNA and disease. J Pathol. 2012 Jan 1;226(2):274–86.
81. Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. Nat Rev Genet. 2005;6(5):389–402.
82. Smith AL, Whitehall JC, Bradshaw C, Gay D, Robertson F, Blain AP, et al. Age-associated mitochondrial DNA mutations cause metabolic remodelling that contributes to accelerated intestinal tumorigenesis. Nat cancer. 2020/09/21. 2020 Oct;1(10):976–89.
83. Satoh M, Kuroiwa T. Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Exp Cell Res. 1991;196(1):137–40.
84. García�Rodríguez LJBT-M in CB. Appendix 1. Basic Properties of Mitochondria. In: Mitochondria, 2nd Edition. Academic Press; 2007. p. 809–12.
85. Taylor RW, Barron MJ, Borthwick GM, Gospel A, Chinnery PF, Samuels DC, et al. Mitochondrial DNA mutations in human colonic crypt stem cells. J Clin Invest. 2003 Nov;112(9):1351–60.
86. Williams SL, Mash DC, Züchner S, Moraes CT. Somatic mtDNA Mutation Spectra in the Aging Human Putamen. PLOS Genet. 2013 Dec 5;9(12):e1003990.
87. Greaves LC, Nooteboom M, Elson JL, Tuppen HAL, Taylor GA, Commane DM, et al. Clonal Expansion of Early to Mid-Life Mitochondrial DNA Point Mutations Drives Mitochondrial Dysfunction during Human Ageing. PLOS Genet. 2014 Sep 18;10(9):e1004620.
88. Steele CD, Abbasi A, Islam SMA, Bowes AL, Khandekar A, Haase K, et al. Signatures of copy number alterations in human cancer. Nature. 2022;606(7916):984–91.
89. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578(7793):112–21.
90. Sugita I, Matsuyama S, Dobashi H, Komura D, Ishikawa S. Viola: a structural variant signature extractor with user-defined classifications. Bioinformatics. 2022 Jan 15;38(2):540–2.
91. Funnell T, Zhang AW, Grewal D, McKinney S, Bashashati A, Wang YK, et al. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. PLOS Comput Biol. 2019 Feb 22;15(2):e1006799.
92. Cameron DL, Baber J, Shale C, Valle-Inclan JE, Besselink N, van Hoeck A, et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. 2021;22(1):202.
93. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177:821-836.e16.
94. Degasperi A, Zou X, Dias Amarante T, Martinez-Martinez A, Koh GCC, Dias JML, et al. Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. Science (80-). 2022 Jul 21;376(6591):abl9283.
95. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149(5):979–93.
96. Maura F, Degasperi A, Nadeu F, Leongamornlert D, Davies H, Moore L, et al. A practical guide for mutational signature analysis in hematological malignancies. Nat Commun. 2019;10(1):2969.
97. Degasperi A, Amarante TD, Czarnecki J, Shooter S, Zou X, Glodzik D, et al. A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. Nat Cancer. 2020;1(2):249–63.
98. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence MS, et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature. 2015;518:360–4.
99. Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science (80-). 2019 Jun 28;364(6447):eaaw2872.
100. Gonzalez-perez A, Sabarinathan R, Lopez-bigas N. Review Local Determinants of the Mutational Landscape of the Human Genome. Cell. 2019;177(1):101–14.
101. Tang B, Pan Z, Yin K, Khateeb A. Recent Advances of Deep Learning in Bioinformatics and Computational Biology. Vol. 10, Frontiers in Genetics. 2019.
102. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E, Al-Shahrour F, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun. 2020;11(1):728.
103. Danyi A, Jager M, de Ridder J. Cancer Type Classification in Liquid Biopsies Based on Sparse Mutational Profiles Enabled through Data Augmentation and Integration. Vol. 12, Life . 2022.
104. Nguyen L, Van Hoeck A, Cuppen E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. Nat Commun. 2022;13(1):4013.
105. Wan JCM, Stephens D, Luo L, White JR, Stewart CM, Rousseau B, et al. Genome-wide mutational signatures in low-coverage whole genome sequencing of cell-free DNA. Nat Commun. 2022;13(1):4953.
106. de Kanter JK, Peci F, Bertrums E, Rosendahl Huber A, van Leeuwen A, van Roosmalen MJ, et al. Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. Cell Stem Cell. 2021;
107. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med. 2018;50(8):1–8.
108. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature. 2020;578(7793):102–11.
109. Elliott K, Larsson E. Non-coding driver mutations in human cancer. Nat Rev Cancer. 2021;21(8):500–9.
110. French JD, Edwards SL. The Role of Noncoding Variants in Heritable Disease. Trends Genet. 2020;36(11):880–91.
111. Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat Genet. 2019;51(6):973–80.
112. Villar D, Frost S, Deloukas P, Tinker A. The contribution of non-coding regulatory elements to cardiovascular disease. Open Biol. 2022 Jul 21;10(7):200088.
113. Frydas A, Wauters E, van der Zee J, Van Broeckhoven C. Uncovering the impact of noncoding variants in neurodegenerative brain diseases. Trends Genet. 2022;38(3):258–72.
114. Bujold D, Morais DA de L, Gauthier C, Côté C, Caron M, Kwan T, et al. The International Human

Epigenome Consortium Data Portal. Cell Syst. 2016 Nov 23;3(5):496-499.e2.
115. Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Ai R, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583(7818):699–710.
116. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan 8;47(D1):D886–94.
117. Farvid MS, Sidahmed E, Spence ND, Mante Angua K, Rosner BA, Barnett JB. Consumption of red meat and processed meat and cancer incidence: a systematic review and meta-analysis of prospective studies. Eur J Epidemiol. 2021;36(9):937–51.
118. McGee EE, Jackson SS, Petrick JL, Van Dyke AL, Adami H-O, Albanes D, et al. Smoking, Alcohol, and Biliary Tract Cancer Risk: A Pooling Project of 26 Prospective Studies. JNCI J Natl Cancer Inst. 2019 Dec 1;111(12):1263–78.
119. van Zandwijk N, Reid G, Frank AL. Asbestos-related cancers: the 'Hidden Killer' remains a global threat. Expert Rev Anticancer Ther. 2020 Apr 2;20(4):271–8.
120. Deng Y, Wang M, Tian T, Lin S, Xu P, Zhou L, et al. The Effect of Hexavalent Chromium on the Incidence and Mortality of Human Cancers: A Meta-Analysis Based on Published Epidemiological Cohort Studies. Vol. 9, Frontiers in Oncology. 2019.
121. Chipperfield MP, Dhomse SS, Feng W, McKenzie RL, Velders GJM, Pyle JA. Quantifying the ozone and ultraviolet benefits already achieved by the Montreal Protocol. Nat Commun. 2015;6(1):7233.
122. Petljak M, Chu K, Dananberg A, Bergstrom EN, Morgen P von, Alexandrov LB, et al. The APOBEC3A deaminase drives episodic mutagenesis in cancer cells. bioRxiv. 2021 Jan 1;2021.02.14.431145.

&

Nederlandse samenvatting
Author contributions by chapter
List of publications
Dankwoord
Curriculum Vitae

Nederlandse samenvatting

DNA en mutaties
Bijna iedere cel in ons lichaam bevat DNA. DNA-moleculen bestaan uit twee suiker-fosfaat ruggengraten die in een helix vorm in elkaar zijn gedraaid. Deze twee ruggengraten zitten aan elkaar vast via twee complementaire nucleotide basenparen, adenine en thymine, en cytosine en guanine. De volgorde van deze nucleotiden vormt een code, die de instructies bevat voor het maken van eiwitten. Deze eiwitten vormen de cellulaire machines en een deel van de bouwstenen die ervoor zorgen dat het menselijke lichaam werkt. In totaal bevat het DNA rond de 3.2 miljard basenparen verdeeld over 22 reguliere chromosomen en 2 geslachtschromosomen. Bij de bevruchting van de eicel komt een set van 23 chromosomen van de vader en een set van 23 chromosomen van de moeder.

Herhaaldelijke celdelingen van de bevruchte eicellen en diens dochtercellen kunnen uiteindelijk leiden tot de ontwikkeling van een menselijk lichaam. Tijdens ieder van deze celdelingen wordt het volledige DNA gekopieerd. Hoewel dit grotendeels goed gaat, kunnen er bij het kopiëren foutjes optreden. Daarnaast is het ook mogelijk dat het DNA beschadigd wordt tussen celdelingen in. Deze beschadigingen kunnen veroorzaakt worden door externe factoren zoals UV-straling, maar ook door interne cel processen. In totaal wordt het DNA per cel per dag ongeveer 10,000 keer beschadigd, maar de meeste beschadigingen worden gerepareerd met behulp van DNA-reparatie eiwitten. Een voorbeeld van DNA-schade is de spontane deaminatie van een cytosine, waardoor het verandert in een thymine. Dit zorgt ervoor dat een thymine gebonden is aan een guanine. Deze combinatie, die normaal niet hoort te gebeuren, kan door een cel opgemerkt en gerepareerd kan worden. Echter als dit niet gerepareerd wordt voordat er een celdeling, plaatsvindt dan wordt er een adenine tegenover de thymine geplaatst en kan deze verandering niet meer gedetecteerd en gerepareerd worden. Dit soort veranderingen in het DNA die tijdens het leven plaatsvinden noemen we somatische mutaties. De combinatie van verschillende soorten DNA-schade met incomplete of foutieve DNA-reparatie zorgt voor verschillende soorten somatische mutaties. Door te kijken naar de verhoudingen tussen verschillende typen mutaties in een cel, is het mogelijk om de processen die de mutaties hebben veroorzaakt te onderzoeken.

Bij een celdeling wordt een somatische mutatie doorgegeven aan de dochtercellen, die het vervolgens weer aan hun eigen dochtercellen doorgeven. Doordat somatische mutaties er constant bijkomen en meestal niet verdwijnen neemt het aantal basis substituties, mutaties waarbij een nucleotide is vervangen door een andere nucleotide, lineair toe met de leeftijd. De snelheid van mutatie accumulatie verschilt tussen organen. Bloed stamcellen krijgen er gemiddeld 15 base substituties per jaar bij, terwijl stamcellen in de dunne darm er meer dan 40 substituties per jaar bij krijgen.

De meeste mutaties hebben geen of nauwelijks effect. Sommige mutaties zorgen er echter voor dat de functie of activiteit van een eiwit verandert. Het is geen probleem als een enkele cel door een mutatie minder goed functioneert of zelfs doodgaat, omdat ons lichaam nog genoeg andere cellen overheeft. Sommige mutaties zorgen er echter voor dat een cel ongecontroleerd gaat delen, of het immuunsysteem kan ontwijken. Meerdere van deze mutaties samen kunnen uiteindelijk leiden tot kanker.

Het onderzoek beschreven in dit proefschrift
Kanker ontstaat in de meeste gevallen door somatische mutaties in een normale cel. Het doel van het onderzoek in dit proefschrift was om technieken te ontwikkelen om somatische mutaties te ontdekken die eerst nog niet makkelijk ontdekt konden worden en om technieken voor het detecteren van patronen in somatische mutaties te verbeteren, zodat we meer inzicht krijgen in de processen die mutaties veroorzaken.

Een manier om de link tussen somatische mutaties en kanker te onderzoeken is om individuen met een verhoogd risico te bestuderen. Jonge kinderen met Downsyndroom hebben een 400 keer zo hoge kans op het krijgen van Downsyndroom geassocieerde acute megakaryoblastische leukemie als reguliere kinderen. Sommige van de mutaties die leiden tot deze vorm van kanker ontstaan al tijdens de foetale ontwikkeling. Om dit te onderzoeken hebben we bloed en darm stamcellen van foetussen met en zonder Downsyndroom geanalyseerd. Door de genomen van losse cellen met elkaar te vergelijken konden we somatische mutaties vinden die tijdens de ontwikkeling hadden plaatsgevonden. Doordat sommige mutaties gedeeld waren tussen cellen konden we ook stambomen maken die lieten zien hoe de cellen aan elkaar gerelateerd waren. We vonden dat mutatie snelheid van cellen tijdens de ontwikkeling 5,8 keer hoger was dan de snelheid na de geboorte. Dit zou kunnen bijdragen aan het hogere risico voor kinderen om leukemie te krijgen vergeleken met volwassenen. Daarnaast vonden we in de stamcellen van foetussen met Downsyndroom nog 34 extra mutaties. Dit aantal extra mutaties lijkt te klein om de verhoogde kans op leukemie in kinderen met Downsyndroom volledig te verklaren, maar kan hier wel aan bijdragen. Andere factoren zoals veranderingen in de activiteit van verschillende genen spelen waarschijnlijk ook een rol. Vervolgens vergeleken we de mutaties in de foetussen met de somatisch mutaties in leukemie monsters. De mutaties in beide groepen werden veroorzaakt door dezelfde mutatieprocessen, wat betekend dat er dus geen extra mutatieprocessen nodig zijn voor het ontstaan van leukemie. De gevonden processen zijn actief in normale cellen gedurende het leven.

Om mutatieprocessen beter te kunnen bestuderen hebben we vervolgens de tweede versie van het “MutationalPatterns” softwarepakket geschreven. De eerste versie van het pakket was gefocust op substituties van een nucleotide door een andere nucleotide. De tweede versie kan echter ook dubbele substituties van twee nucleotiden naast elkaar analyseren. Inserties en deleties (indels), waarbij er een of een aantal nucleotiden aan het DNA worden toegevoegd of verwijderd kunnen ook worden geanalyseerd. We hebben ook functies toegevoegd die de mutatieprocessen die actief zijn geweest in een monster kunnen ontdekken met minder vals positieven. Daarnaast hebben we functies toegevoegd die mutatieprocessen kunnen onderzoeken die alleen actief zijn in een deel van het genoom. De nieuwe versie van MutationalPatterns hebben we vervolgens getest op cellijnen waar specifieke genen die belangrijk zijn voor het repareren van DNA waren uitgeschakeld. De mutatie patronen die we vonden in deze cellijnen kwamen overeen met wat we verwachtten.

Vervolgens hebben we gekeken naar somatische mutaties in mitochondriën. Mitochondriën zijn organellen die verantwoordelijk zijn voor de energiehuishouding van een cel. Ze hebben hun eigen circulaire DNA, waarvan er meerdere kopieën aanwezig zijn per cel. Dit maakt het lastig om mutaties in het mitochondriale DNA (mtDNA) te ontdekken, wat dan ook niet standaard wordt gedaan. Door gebruik te maken van specifieke software en streng vals positieve varianten te filteren konden we mutaties detecteren in mitochondriën. We vonden dat het aantal somatische mutaties in mitochondriën van normale cellen lineair toeneemt met de leeftijd, net zoals in de celkern waar de rest van het DNA zich bevindt. Het aantal kopieën van het mtDNA per cel veranderde niet met de leeftijd, maar verschilde wel tussen het bloed, de dikke darm, en de dunne darm. Vervolgens vergeleken we het aantal mutaties in mitochondriën van normale cellen met kanker. Uit onze vergelijking bleek dat de meerderheid van de mitochondriale mutaties in kanker monsters al hadden plaatsgevonden voor het ontstaan van de kanker, toen de cel nog normaal was. Het mutatieproces dat de mitochondriale mutaties veroorzaakte was ook hetzelfde tussen de normale- en kankercellen. Vervolgens hebben we gekeken naar cellen die in het lab zijn behandeld met chemotherapie en andere kankerbehandelingen. Deze behandelingen leken geen groot effect te hebben op het mitochondriaal genoom.

De hierboven beschreven experimenten in normale cellen zijn gedaan door losse cellen te kweken in het laboratorium, zodat we genoeg DNA hebben om het genoom te analyseren. Dit werkt echter alleen voor stamcellen en niet voor het grote aantal “gedifferentieerde” cellen die minder kunnen delen. Om het genoom van deze cellen toch te kunnen analyseren is recent de PTA-techniek ontwikkeld, die het DNA van een cel artificieel kopieert zodat er genoeg materiaal is voor een analyse. Deze methode is een stuk sensitiever en accurater dan eerdere methodes, maar resulteert alsnog in een flink aantal vals positieve artefacten. We hebben de “PTA analysis toolkit” (PTATO) ontwikkeld om deze vals positieve artefacten te onderscheiden van echte mutaties voor substituties, indels, en structurele varianten waarbij er een groter deel van het genoom is geamplificeerd of verwijderd. PTATO doet dit door onder andere te kijken naar de patronen van potentiële mutaties en door te kijken of potentiële mutaties op het DNA van of de vader of de moeder zitten. Als een mutatie op zowel het DNA van de moeder als de vader wordt aangetroffen, dan is het waarschijnlijk een artefact, omdat de kans zeer klein is dat twee dezelfde mutaties op exact dezelfde locatie hebben plaatsgevonden. We hebben PTATO gevalideerd op zowel onze eigen data als op een externe dataset. Uiteindelijk hebben we PTATO toegepast op monsters van een leukemiepatiënt. We vonden dat sommige van de mutaties die de kanker veroorzaakten al jaren voor de diagnose hadden plaatsgevonden. Daarnaast bleek het dat kankercellen nog steeds konden differentiëren.

In dit proefschrift hebben we gekeken naar somatische mutaties in normale cellen en de processen waardoor ze worden veroorzaakt. Hiervoor hebben we ook softwarepakketten geschreven die gebruikt kunnen worden in verdere onderzoeken. Hiermee kunnen we het ontstaan van kanker beter begrijpen wat uiteindelijk zou kunnen leiden tot een betere behandeling of zelfs preventie.

Author contributions by chapter

Chapter 1: Introduction: The dynamics of somatic mutagenesis during life in humans
FM, RvB, and SM jointly wrote the manuscript. RvB and SM supervised the project. FM adapted the manuscript for this thesis.

Chapter 2: Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis
KALH, MV, and SMCSL performed sample isolation. KALH and TP performed fluorescence-activated cell sorting (FACS). KALH, MV, and EK performed clonal expansions and supervised sequencing. KALH, FM, and RB wrote the manuscript. FM performed bioinformatic analyses. SMCSL and MLH collected fetal material. RB designed and supervised the study. All authors reviewed the manuscript.

Chapter 3: MutationalPatterns: the one stop shop for the analysis of mutational processes
FM, RvB, and AMB wrote the manuscript. FM and JdK developed and implemented the package. FM and RO maintain the package. AMB and MV generated the data. FM and MJvR analyzed the data. AvH, BvdR, and EC tested the package and provided feedback. All authors read and approved the final manuscript.

Chapter 4: Mutation accumulation in mitochondrial DNA of cancers resembles mutagenesis in normal stem cells
FM and JTD gathered the data and performed bioinformatic analyses. FM and RB wrote the manuscript. RB designed and supervised the study.

Chapter 5: Investigation of single-cell genomes at nucleotide resolution using the PTA Analysis Toolkit (PTATO)
FM, SM, and RvB jointly wrote the manuscript. FM, SM, and MJvR wrote PTATO. FM, SM, MJvR, and JdK performed bioinformatic analyses. SM, EB, IvdW, EA, DR, NMG, MV, and AMB generated the samples. RvB supervised the study.

Chapter 6: General discussion
FM wrote the discussion.

List of publications

Manders, F., van Dinter J, van Boxtel R. *Mutation accumulation in mtDNA of cancers resembles mutagenesis in normal stem cells*. iScience. 2022; 25(12):105610

Manders, F., Brandsma AM, de Kanter J, Verheul M, Oka R, van Roosmalen MJ, et al. *MutationalPatterns: the one stop shop for the analysis of mutational processes*. BMC Genomics. 2022;23(1):134.

Manders. F., van Boxtel R, Middelkamp S. *The Dynamics of Somatic Mutagenesis During Life in Humans*. Vol. 2, Frontiers in Aging. 2021

Manders, F.*, Hasaart, K.A.L.*, van der Hoorn, M. et al. *Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis*. Sci Rep 2020;10(1):12991.

Yu Y, **Manders F**., Grinwis GCM, Groenen MAM, Crooijmans RPMA. *A recurrent somatic missense mutation in GNAS gene identified in familial thyroid follicular cell carcinomas in German longhaired pointer dogs*. BMC Genomics. 2022;23(1):669.

Meister MT, Groot Koerkamp MJA, de Souza T, Breunis WB, Frazer-Mendelewska E, Brok M, DeMartino, J, **Manders, F**., et al. *Mesenchymal tumor organoid models recapitulate rhabdomyosarcoma subtypes*. EMBO Mol Med. 2022 Aug 2;n/a(n/a):e16001.

Hasaart KAL, **Manders. F.**, Ubels J, Verheul M, van Roosmalen MJ, Groenen NM, et al. *Human induced pluripotent stem cells display a similar mutation burden as embryonic pluripotent cells in vivo*. iScience. 2022;25(2):103736.

Brandsma AM, Bertrums EJM, van Roosmalen MJ, Hofman DA, Oka R, Verheul M, **Manders, F.**, et al. *Mutation Signatures of Pediatric Acute Myeloid Leukemia and Normal Blood Progenitors Associated with Differential Patient Outcomes*. Blood Cancer Discov. 2021 Sep 1;2(5):484 LP – 499.

de Kanter JK, Peci F, Bertrums E, Rosendahl Huber A, van Leeuwen A, van Roosmalen MJ, **Manders, F.**, et al. *Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients*. Cell Stem Cell. 2021.

Hasaart KAL, Bertrums EJM, **Manders, F.**, Goemans BF, van Boxtel R. *Increased risk of leukaemia in children with Down syndrome: a somatic evolutionary view*. Expert Rev Mol Med. 2021/04/27. 2021;23:e5.

Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A. van Hoeck, A., Wood, H.M., Nomburg, Jason., Gurjao, C., **Manders, F.**, et al. *Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli*. Nature 580, 269–273 2020;580:269–73.

Huber, A. R., **Manders, F.**, Oka, R., van Boxtel, R. *Characterizing Mutational Load and Clonal Composition of Human Blood*. J. Vis. Exp. 2019 Jul;(149).

Dankwoord

Na vier jaar is mijn PhD ineens voorbij. (Vooral het laatste jaar ging opeens hard). Het werk beschreven in dit proefschrift was niet mogelijk geweest zonder de hulp van een hoop mensen, die ik hier graag wil bedanken.

Ruben, vier jaar geleden heb je me aangenomen als PhD-student. Jouw enthousiasme was erg aanstekelijk en hielp me wanneer een project tegenzat of wanneer ik (te) pessimistisch was over resultaten. Je kwam vaak vol enthousiasme het kantoor binnen met ideeën voor projecten. De vrijheid die je me gaf om mijn projecten in te vullen was erg fijn. Daarnaast waardeerde ik je feedback ook altijd erg.

Frank, ik vond de gesprekken die ik met je heb gehad erg fijn. Ondanks dat je ziek werd heb je nog zo lang mogelijk (meer dan) je taken als promotor uitgevoerd. We hebben zelfs nog een keer gebeld over het vervolg van mijn carrière, toen je niet naar het PMC kon komen. Het is erg jammer dat je door je ziekte mogelijk niet bij het einde kan zijn en ik hoop dat het snel weer wat beter gaat.

Edwin en **Patrick**, tijdens de commissie-meetings hadden jullie altijd goede kritische vragen en suggesties, die mijn promotie traject op het goede spoor hielden. Daarnaast waren de bevestigingen dat ik op schema lag ook erg fijn. Edwin, bedankt dat je Franks taak als promotor wou overnemen, ook al ben je gestopt met je groep op het UMCU. Ook wil ik de overige leden van mijn leescommissie: **Gerard Kops**, J**osé Borghans**, **Berend Snel,** en **Lodewijk Wessels**, bedanken voor hun tijd en voor het doornemen van mijn proefschrift.

Jurrian, bedankt voor de discussies en de adviezen je me hebt gegeven. Ik vond het leuk om samen met je te werken en om samen te boulderen. Bedankt dat je mijn paranimf wilt zijn tijdens de verdediging.
Wijnand, samen boulderen, mario karten, of nerden was altijd een fijne manier om te ontspannen. Daarnaast was het ook altijd fijn om te kunnen praten als er iets wat minder ging. Ook jij bedankt dat je mijn paranimf wilt zijn.

Ik wil ook alle boxtel-buddies bedanken voor hun hulp en gezelligheid. Vier jaar geleden waren we een kleine groep, maar dit veranderde al snel. Groepsactiviteiten waren altijd erg gezellig met als afsluiting nog een mooie retreat op mijn laatste (officiële) dag.
Eline, ik vond het leuk om samen te werken aan de RUNX1-RUNX1T AML. Nog een paar maanden en dan is jouw verdediging al. **Flavia**, good luck with the final chapters. I enjoyed your good italian food and strong opinions.

Mark v R., bedankt voor het beantwoorden van al mijn vragen en sorry voor de keren dat ik de HPC liet overlopen. **Lucca**, we zijn samen begonnen met boulderen wat uiteindelijk is uitgegroeid tot een erg leuke wekelijkse activiteit. Succes met je HR-meetings. **Mark V.** en **Niels**, geen idee wat jullie allemaal in het lab doen, maar het is vast meer werk dan op de entertoets drukken. **Joske**, onze machine-learning expert, bedankt voor al je advies en feedback. **Sjors**, we hebben samen een mini-review geschreven en aan het PTA-project gewerkt. Ik vond het fijn om dit samen te doen. Ga daarnaast vooral door met je memes. **Annemarie**, bedankt voor het georganiseerd houden van de groep en me een duwtje in de rug geven toen het nodig was. **Vera**, het moet lastig zijn geweest om van groep te veranderen tijdens je PhD, maar ik vond het erg gezellig dat je bij ons kwam. **Inge**, gezellig dat je tot twee keer toe bij onze groep kwam. **Rico**, het was fijn om een extra bioinformaticus in de groep te hebben. Zeker eentje met lekker droge humor. **Diego**, tu eres una buena adición al grupo. Te voy a ver en la pista de baile. **Laurianne**, I'm impressed you joined bouldering even though you have a fear of heights. **Alexander**, you started when I was almost gone already. Good luck with your PhD and I hope you enjoy my spot. **Anaïs**, leuk dat je weer terug bent bij de vanBoxtel groep. Succes met je PhD.
All former Boxtel-buddies: **Melissa**, **Niels**, **Susanna**, **Madalena**, **Sophie G**, **Sophie**, **Kees**, **Marta**, **Damon**, **Andrea**, **Annina**, thanks for making the lab such a fun place.
Axel, ik heb altijd genoten van het voetballen en squashen. Alleen met het klimmen ben ik helaas te laat begonnen om mee te doen. Bedankt dat ik je paranimf mocht zijn en veel succes met de rest van je postdoc in Spanje. **Karlijn**, we hebben samen mijn eerste paper geschreven. Alhoewel ik soms koppig kon zijn, was het volgens mij wel een goede samenwerking. **Rurika**, you helped me a lot during the beginning of my PhD. Thank you for that. **Arianne**, bedankt voor je adviezen en fijn dat ik jouw knock-out data kon gebruiken. **Miriam**, succes met je eigen groep. **Jip**, ik hoop dat je veel geleerd hebt tijdens je stage. Ik heb in ieder geval wel veel geleerd (Cliché, maar waar). Het mitochondriën project dat jij bent begonnen is een hoofdstuk in mijn boekje geworden en zelfs gepubliceerd. Leuk dat je in het PMC bent blijven plakken.

Axel, **Jurrian**, **Joske**, en **Kees**. Ik vond het erg gezellig om tijdens de lockdowns bij elkaar thuis te werken en op die manier toch nog wat sociaal contact te hebben. Bedankt voor de lekkere lunches.

Lucca, **Jurrian**, **Niels**, **Laurianne**, and **Alexander**. Thanks for joining the bouldering. I always really enjoy it and I hope I will see you there again.

Michael, bedankt voor alle leuke gesprekken over fantasy en games. Ik vond het erg leuk om bij je huwelijk te zijn en wens je nog veel geluk toe samen met Julian.

Dilys, ik vond het altijd erg gezellig om samen te eten of wat thee te drinken. Succes met de laatste loodjes. Additionally, I want to thank all the other people at the Máxima for great discussions and fun times during the borrels.

Eline, **Leiah**, **Nienke**, **Madeleine**, **Thomas**, **Yvette**, **Maroussia**, **Anne**, **Sruthi**, **Irene**, thanks for the nice times in the PriMa PhD group.

Wim, **Axel**, **Moritz**, thanks for organizing the football. I always enjoyed it even though I was pretty bad. Also thank you to all the other players: **Saman**, **Lars**, **Margit**, **Muhammad**, **Winnie**, **Enric**, **Erik**, **Javi**, **Jan**, **Lucas**, **Nico**, **Cayetano**, **Jens**, **Moritz**, **Kostas**, **Mike**, **Joost**, **Louk**, **Guy**.

Jeroen en **Max**, samen een bordspel doen, film kijken, of een cocktail maken was altijd erg gezellig en een leuke afleiding van werk.
Wijnand, **Aiko**, **Cathrin**, **Inge**, **Laura**, en **Kevin**, het was altijd leuk om gezellig samen wat te doen.
Alexander en **Joeri**, het was altijd fijn om over dingen te kunnen praten en om gewoon lekker wat af te spreken. We zijn al lang bevriend en ik hoop dat dit nog lang zo blijft.

Dirk, ik zie je niet meer zo vaak nu je in Engeland zit, maar ik vind het altijd fijn om te bellen. Samen Londen bekijken was erg leuk.
Wout, ik vind het altijd leuk om te praten over series of om samen een bordspel te doen. Veel succes met de rest van de PABO.
Lieve **papa** en **mama**, bedankt voor alle steun. Jullie hebben me altijd aangemoedigd om te doen wat ik leuk vind en hebben me altijd ondersteund als het wat minder ging en daarvoor ben ik jullie erg dankbaar.

Curriculum vitae

Freek Martijn Manders was born on 3 January 1995 in Veghel, the Netherlands. In 2013 he obtained his high school degree at "Stedelijk Gymnasium Nijmegen", after which he started his bachelor medical biology at the Radboud University in Nijmegen. During this time he performed a combined internship / Honours academy project in the lab of Dr. Colin Logie, where he used 4C-sequencing and bioinformatics to study topologically associated domains. In 2016 he started his master in medical biology (track medical epigenomics), also at the Radboud University. He performed his first internship in the group of Prof. Dr. Michiel Vermeulen at the Radboud Institute for Molecular Life Sciences, where he used mass-spectrometry and bioinformatics to investigate ubiquitin interactions in the nucleus. He performed his second internship in the group of Prof. Dr. Edwin Cuppen at the University Medical Center Utrecht, where he investigated de novo SNVs in patients with neurodevelopmental disorders. In September 2018 he started his PhD in the group of Dr. Ruben van Boxtel at the Princess Máxima Center for pediatric oncology. During his PhD he improved, developed, and applied bioinformatic methods to identify somatic mutations and characterize patterns within them. The results of his work are described in this thesis.