Mutation accumulation in mtDNA of cancers resembles mutagenesis in normal stem cells

Freek Manders, Jip van Dinter, Ruben van Boxtel
r.vanboxtel@prinsesmaximacentrum.nl

Highlights
Mitochondria in normal cells accumulate mutations with age
Most mitochondrial mutations in cancer result from healthy mutation accumulation
Cancer therapies do not result in many mitochondrial mutations
Mutation accumulation in mtDNA of cancers resembles mutagenesis in normal stem cells

Freek Manders, Jip van Dinter, and Ruben van Boxtel

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 aging and cancer. 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 patients with leukemia have an increased mitochondrial mutation load. Finally, we show that secondary pediatric cancers and chemotherapy treatments do not impact the mitochondrial mutation load and mtDNA 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.

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. A single cell harbors many mitochondria, ranging from a couple dozen to more than a thousand depending on the cell type. Mitochondria contain their own mtDNA, of which up to 15 copies can be present per mitochondrion. The mtDNA is circular and, even though it is only 16.6 kb, 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. Genetic variation in the mtDNA of a cell is often present in only a subset of its mitochondria, a phenomenon known as heteroplasmy.

Defects in the mitochondrial genome have been associated with the development of a variety of neurodegenerative diseases as well as aging. In addition, mutations in mtDNA have also been suspected to play a role in the onset or progression of cancer. 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. 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. 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. 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. 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

Cataloging somatic mitochondrial mutations

A mitochondria-specific sequencing analysis pipeline was recently developed as a part of the genome analysis toolkit. 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 (Figure 1, Table S1). 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 (Figure S1A).

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–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 gray background shows the 95% confidence interval of the model, whereas the light gray 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 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) (Figures S1B and S1C, Table S2). 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 genome30 (Figures S1D and S1E). 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 (Figure S1 F). Most of the genic substitutions were predicted to have a low to moderate effect, suggesting that they are unlikely to have a large physiological effect (Figure S1 G).

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; Figure 1B). The mutation load of HSPCs did have a weak correlation between mtDNA copy numbers and age (p = 0.0451; generalized linear model; Figure 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; χ² = 0.313; Chi-squared test; Figure S1). 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; Figures 1C and 1D). 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 show that mitochondria in stem cells gradually accumulate mutations with age in multiple tissues at comparable rates.

mtDNA 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; Figure 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 patients with cancer.12 The depth of sequencing did not influence our results, as WGS samples sequenced at 15X and 30X had comparable copy numbers (Figure S2). However, the mitochondrial copy numbers did differ between the various cell types (Figure 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.13 This observation indicates that the variation in mtDNA copy numbers between these cancers is not necessarily caused by mitochondrial dysfunction due to the malignant phenotype, but likely reflects the differences already found between healthy tissues of which these cancers arise.

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
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. Additionally, we analyzed the mitochondrial DNA 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 were included in the PCAWG consortium. After combining these 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; Figure 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 mitochondrial 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 was a synonymous base substitution, was present in both a normal HSPC and blood cancer sample, indicating that the mutations we observed are random passengers and not recurrent drivers. The spectra of the mitochondrial DNA mutations in normal HSPCs consisted of mostly C>T substitutions on the heavy strand and T>C substitutions on the light strand (Figure 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 (Figure 3C). Additionally, mutations in both the normal HSPCs and the blood cancers were distributed across the mitochondrial genome (Figure S3A). These observations suggest that the mitochondrial DNA substitutions found in blood cancer samples are caused by the same mutational processes, likely related to mitochondrial DNA replication, as the substitutions found in normal blood, which further underlines the idea that most mitochondrial mutations in blood cancer are the result of normal age-related mutagenesis. Interestingly, we observed a slightly higher ratio of missense substitutions in the normal HSPCs compared to the blood cancers (p = 0.048; Chi-squared test; Figure 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; 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; Chi-squared test), as was the distribution of
mutations across the mitochondrial genome (Figure 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; Figure 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 (Figures 3E and 3F, 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 \( p = 0.0005 \) (\( \chi^2 = 23.477 \); Chi-squared test). However, the trinucleotide profiles of the C>T mutations on the light strand are quite similar (Figure 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. In line with this, the ratio of missense mutations was similar between colon cancer and normal colon stem cells (\( p = 0.518; \chi^2 = 0.6 \); Chi-squared test; Figure S3E), and mutations in both groups were distributed across the mitochondrial genome (Figure 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 (\( p = 0.464; \chi^2 = 0.6 \); Chi-squared test) as was the distribution of mutations across the mitochondrial genome (Figure 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 (\( p < 0.0001; p = 0.0001 \); linear mixed-effects model; Figures S3H and S3I). 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 (\( p < 0.0001; W = 149 \); Wilcoxon rank-sum test; Figure 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; Figure 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 mutation 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 (Figure S4A). However, one sample was a statistical outlier, with an 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; Figure 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
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 patients with leukemia (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 value shows 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 = Doxorubicin (2 mutations; 5 samples; 2 donors), MAPH = Maphosphamide (0 mutations; 6 samples; 2 donors), RAD = X-ray (2 mutations; 9 samples; 4 donors), VINCRIS = Vincristine (0 mutations; 6 samples; 2 donors), GCV = Ganciclovir (0 mutations; 3 samples; 1 donor), FC = Foscarnet (0 mutations; 2 samples; 1 donor). The p value shows the significance of the difference in the number of base substitutions per clone between treatments (one-way ANOVA).

(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 = Doxorubicin (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). The p value shows the significance of the difference in the mtDNA copy numbers per clone between treatments (one-way ANOVA).
To investigate whether this also holds true for mtDNA, we analyzed mitochondrial mutation 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; Figure 4B). Secondary leukemias also did not contain an increased number of mitochondrial substitutions compared to primary leukemias from the same patient (Figure 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.

Secondary leukemias did not have a significantly different mtDNA copy number than HSPCs from healthy donors (Figure S4A). However, similar to the normal HSPCs, one sample was a statistical outlier with an 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. 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; Figure 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; Figure 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. One possible explanation for the limited effect of cancer and its treatments is that the mtDNA damage they cause might be resolved by cells clearing their damaged mitochondria. While our data suggest that damage to the mitochondrial genome does not play a major role in the development of cancer, some individual mitochondrial mutations can still contribute to cancer development. The increased mutation loads in pediatric cancer patients were 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 suggest 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.

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
Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Read alignment and variant calling
  - Somatic variant filtering
  - Mutation load accumulation
  - Comparison mitochondrial and nuclear mutation load
  - Copy number analysis
  - Mutation spectra
  - Comparisons to PCAWG
  - Gene mutations
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION
Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105610.

ACKNOWLEDGMENTS
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. Additionally, we want to thank Joske Ubels for providing feedback on the manuscript. Finally, we want to thank Axel R. Huber, Jurrian de Kanter, Flavia Peci, Arianne M. Brandsma, Runika 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
The authors declare no competing interests.

Received: June 2, 2022
Revised: October 26, 2022
Accepted: November 14, 2022
Published: December 22, 2022
REFERENCES

1. Hengartner, M.O. (2000). The biochemistry of apoptosis. Nature 403, 770–776. https://doi.org/10.1038/35037710.

2. Duchen, M.R. (2000). Mitochondria and calcium: from cell signalling to cell death. J. Physiol. 529, 57–68. https://doi.org/10.1111/j.1469-7793.2000.00057.x.

3. Zong, W.-X., Rabinowitz, J.D., and White, E. (2016). Mitochondria and cancer. Mol. Cell 61, 667–676. https://doi.org/10.1016/j.molcel.2016.02.011.

4. Schon, E.A., DiMauro, S., and Hirano, M. (2012). Human mitochondrial DNA: roles of inherited and somatic mutations. Nat. Rev. Genet. 13, 878–890. https://doi.org/10.1038/nrg3275.

5. Garcia-Rodriguez, L.J. (2007). Appendix 1. Basic properties of mitochondria. Methods Cell Biol. 809–812. https://doi.org/10.1016/S0076-6846(06)80040-3.

6. Satoh, M., and Kuroiwa, T. (1991). Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. Exp. Cell Res. 196, 137–140. https://doi.org/10.1016/0014-4827(91)90467-9.

7. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., et al. (1981). Sequence and organization of the human mitochondrial genome. Nature 290, 457–465. https://doi.org/10.1038/290457a0.

8. Stewart, J.B., and Chinery, P.F. (2015). The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat. Rev. Genet. 16, 530–542. https://doi.org/10.1038/nrg3966.

9. Alston, C.L., Rocha, M.C., Lax, N.Z., Turnbull, D.M., and Vincent, A.E. (2020). The rise and rise of mitochondrial DNA mutations. Open Biol. 10, 200061. https://doi.org/10.1098/rsob.200061.

10. Hasasta, K.A.L., Manders, F., van der Hoorn, M.L., Verheul, M., Popolonski, T., Kuijk, E., de Sousa Lopes, S.M.C., and van Boxtel, R. (2020). Mutation accumulation and developmental lineages in normal and Down syndrome human fetal hematopoiesis. Sci. Rep. 10, 12991. https://doi.org/10.1038/s41598-020-69822-1.

11. de Kanter, J.K., Peci, F., Bertrums, E., Rosendahl Huber, A., van Leeuwen, A., van Roosmalen, M.J., Manders, F., Verheul, M., Oka, R., Brandsma, A.M., et al. (2021). Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. Cell Stem Cell 28, 1726–1739.e6. https://doi.org/10.1016/j.stem.2021.07.012.

12. Bolouri, H., Farrar, J.E., Triche, T., Jr., Ries, R.E., Lim, E.L., Alonzo, T.A., Ma, Y., Moore, R., Mungall, A.J., Marra, M.A., et al. (2018). The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat. Med. 24, 103–112. https://doi.org/10.1038/nm.4439.

13. Bertrums, E.J.M., Rosendahl Huber, A.K.M., de Kanter, J.K., Brandsma, A.M., van Leeuwen, A.J.C.N., Verheul, M., van den Heuvel-Eibrink, M.M., Oka, R., van Roosmalen, M.J., de Groot-Kruisen, H.A., et al. (2022). Elevated mutational age in blood of children treated for cancer contributes to therapy-related myeloid neoplasms. Cancer Discov. 12, 1860–1872. https://doi.org/10.1158/2159-8290.CD-22-0120.

14. Blöckl, F., de Ligt, J., Jager, M., Sasselli, V., Roeink, 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. https://doi.org/10.1038/nature19768.

15. Osorio, F.G., Rosendahl Huber, A., Oka, R., Verheul, M., Patel, S.H., Hasasta, K., de la Fonteine, L., Varela, I., Camargo, F.D., and van Boxtel, R. (2018). Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. Cell Rep. 25, 2338–2346.e4. https://doi.org/10.1016/j.celrep.2018.11.014.

16. Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sandau, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., et al. (2019). The landscape of somatic mutation in normal colorectal epithelium at single-nucleotide resolution. Nature 567, 532–537. https://doi.org/10.1038/s41586-019-1672-7.

17. Brandsma, A.M., Bertrums, E.J.M., van Roosmalen, M.J., Hofman, D.A., Oka, R., Verheul, M., Manders, F., Uebels, J., Belderbos, M.E., and van Boxtel, R. (2021). Mutation signatures of pediatric acute myeloid leukemia and normal blood progenitors associated with differential patient outcomes. Blood Cancer Discov. 2, 484–499. https://doi.org/10.1158/2643-3230.BCD-21-0010.

18. Williams, S.L., Mash, D.C., Zúñiger, S., and Moraes, C.T. (2013). Somatic mtDNA mutation spectra in the aging human putamen. PLoS Genet. 9, e1003990.

19. Kennedy, S.R., Salk, J.J., Schmitt, M.W., and Loeb, L.A. (2013). Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet. 9, e1003794. https://doi.org/10.1371/journal.pgen.1003794.

20. Ju, Y.S., Alexandrov, L.B., Gerstung, M., Martincorena, I., Nik-Zamil, S., Ramakrishna, M., Davies, H.R., Papoulil, E., Gumund, G., Shlien, A., et al. (2014). Origins and functional consequences of somatic mitochondrial mutations in human cancer. Elife 3, e03923. https://doi.org/10.7554/eLife.03923.

21. Baker, K.T., Nachman悻, D., Kumar, S., Emmond, J., Maksakli, C., Brentnell, T.A., Kennedy, S.R., and Ruscio, R.A. (2019). Mitochondrial DNA mutations are associated with ulcerative colitis preneoplasia but tend to be negatively selected in cancer. Mol. Cancer Res. 17, 488–498. https://doi.org/10.1158/1541-7786.MCR-18-0520.

22. Greaves, L.C., Reeve, A.K., Taylor, R.W., and Turnbull, D.M. (2012). Mitochondrial DNA and disease. J. Pathol. 226, 274–286. https://doi.org/10.1002/path.3028.

23. Taylor, R.W., and Turnbull, D.M. (2005). Mitochondrial DNA mutations in human disease. Nat. Rev. Genet. 6, 389–402.

24. Yuan, Y., Ju, Y.S., Kim, Y., Li, J., Wang, Y., Yoon, C.J., Yang, Y., Martincorena, I., Creghton, C.J., Weinstein, J.N., et al. (2020). Comprehensive molecular characterization of mitochondrial genomes in human cancers. Nat Genet. 52, 342–352. https://doi.org/10.1038/s41438-019-0557-x.

25. Smith, A.L., Whitehall, J.C., Bradshaw, C., Gay, D., Robertson, F., Blain, A.P., Hudson, G., Pyle, A., Houghton, D., Hunt, M., et al. (2020). Age-associated mitochondrial DNA mutations cause metabolic remodelling that contributes to accelerated intestinal tumorigenesis. Nat. Cancer 1, 976–989. https://doi.org/10.1038/s43018-020-00112-5.
cells. STAR Protoc. 3, 101361. https://doi.org/10.1016/j.xprot.2022.101361.

35. Picard Toolkit (2019). Broad Institute, GitHub Repos http://broadinstitute.github.io/picard/

36. Bonfield, J.K., Marshall, J., Danecbek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R.M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience 10, 772. https://doi.org/10.1093/gigascience/giaa176.

37. Rosendahl Huber, A., van Leeuwen, A.J.C.N., Peci, F., de Kanter, J.K., Bertrums, E.J.M., and van Boxtel, R. (2022). Whole-genome sequencing and mutational analysis of human cord-blood derived stem and progenitor cells. STAR Protoc. 3, 101361. https://doi.org/10.1016/j.xprot.2022.101361.

38. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature 578, 94–101. https://doi.org/10.1038/s41586-020-1943-3.

39. Lott, M.T., Leipzig, J.N., Derbeneva, O., Xie, H.M., Chalkia, D., Sarmady, M., Procaccio, V., and Wallace, D.C. (2013). mtDNA variation and analysis using mitomap and mitomaster. Curr. Protoc. Bioinformatics 44, 1.23.1–1.23.26. https://doi.org/10.1002/0471250953.bob0123s44.

40. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. 48, D682–D688. https://doi.org/10.1093/nar/gkaa966.

41. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). 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. https://doi.org/10.4161/fly.19695.
## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Deposited data      |        |            |
| Whole-genome sequence data from normal colon, intestine and liver stem cells. | (Blokzijl et al., 2016)²⁴ | European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) Accession Numbers EGAS00001001682 and EGAS00001000881 |
| Whole-genome sequence data from normal HSPCs from healthy donors. | (Osorio et al., 2018)¹⁵ | European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) Accession Number EGAS00001003068 |
| Whole-genome sequence data from fetuses with and without Trisomy 21. | (Hasaart et al., 2020)²⁴ | European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) Accession Number EGAS00001003982 |
| Whole-genome sequence data from pediatric AML patients | (Brandsma et al., 2021)¹⁷ | European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) Accession Number EGAS00001004593 |
| Whole-genome sequence data from HSPCs of pediatric patients that underwent a hematopoietic stem cell transplant | (de Kanter et al., 2021)²⁵ | European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) Accession Number EGAS00001004926 |
| Whole-genome sequence data from pediatric leukemia patients and cord blood cells that were treated with chemotherapies and X-ray in-vitro | (Bertrums et al., 2022)²⁷ | European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) Accession Number EGAS00001005141 |
| Whole-genome sequence data from pediatric AML samples | (Bolouri et al., 2018)²⁶ | Database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) Accession Number phs000218 |
| Haplotype specific blacklist v102 | (Lott et al., 2013)³⁹ | https://www.mitomap.org/MITOMAP |
| Mitochondrial gene lengths v104 | (Yates et al., 2020)⁴⁰ | https://www.ensembl.org/index.html |
| Mutation calls and copy numbers | This paper | https://github.com/ProjectsVanBox/mitochondria_mutation_accumulation |
| Software and algorithms |        |            |
| Genome analysis toolkit v4.1.3.0 | (McKenna et al., 2010)²⁸ | https://gatk.broadinstitute.org/hc/en-us |
| Mitolib v0.1.2 | https://github.com/haanss/mitolib | https://github.com/haanss/mitolib |
| Picard SamToFastq v2.24.1 | Picard toolkit’, 2019⁴¹ | https://broadinstitute.github.io/picard/ |
| Picard toolkit | ‘Picard toolkit’, 2019⁴¹ | https://broadinstitute.github.io/picard/ |
| bgzip v1.0 | (Bonfield et al., 2021)⁴² | http://www.htslib.org/doc/bgzip.html |
| NF-IAP v1.2 | UMCU genetics | https://github.com/UMCUGenetics/NF-IAP |
| SnpEff v4.3t | (Cingolani et al., 2012)⁴³ | http://pcingola.github.io/SnpEff/ |
| R language v3.6.3 | (R Core Team, 2018)⁴⁴ | https://www.r-project.org/ |
| nlme v3.1-148 | (Pinheiro et al., 2018)⁴⁵ | https://cran.r-project.org/ |
| lme4 v1.1-23 | (Bates et al., 2015)⁴⁶ | https://cran.r-project.org/ |
| ggeffects v0.15.0 | (Lüdecke, 2018)⁴⁷ | https://cran.r-project.org/ |
| MutationalPatterns v3.3.4 | (Manders et al., 2022)⁴⁸ | https://github.com/ToolsVanBox/MutationalPatterns |
| Modified mitochondrial pipeline | This paper; Genome Analysis Toolkit | https://github.com/ProjectsVanBox/mitochondria_mutation_accumulation; https://github.com/broadinstitute/gatk/tree/master/scripts/mitochondria_m2_wdl |
| Custom analysis scripts | This paper | https://github.com/ProjectsVanBox/mitochondria_mutation_accumulation |
RESOURCE AVAILABILITY

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 and code availability
- Data are available on EGA under accession numbers EGA: EGAS000001001682, EGAS000001000881, EGAS000001003068, EGAS000001003982, EGAS000001004593, EGAS000001004926, EGAS000001005141. The TARGET data are available on the database of Genotypes and Phenotypes (dbGaP) with accession number dbGaP: phs000218. The PCAWG mutation frequencies were provided by Young Seok.
- 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.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All 532 samples used in this study are available on EGA as described in the Data availability statement. An overview of all samples and their source is given in Table S1. 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.13 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.11 The samples from donors MH2, NR1, and NR2 were generated from fetal blood.26 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, PARXLYR, PASDKZ, PASDXR, PASFHK, PASFFJ, PASLZE, PASSLT, PASVJS, PASYWA, PATISD, and PATKJK 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.26 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.25,27,37 Bulk cord blood was treated with the relevant treatment for 3 days, after which a single HSPC was clonally expanded and sequenced.37

The samples that we analyzed with the mitochondrial pipeline were compared to previously identified substitutions in adult cancers from PCAWG.12

METHOD DETAILS

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. This merging step was also done in the original study that generated these samples, but had to be repeated because of the re-alignment to hg38.24

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.49 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. Additionally, the “autosomal-coverage” argument was supplied with the mean nuclear coverage, the “stats” argument was supplied with the merged “stats” file, and the “contamination-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.29,30,51 We modified the pipeline to also flag common variants using haplotype specific blacklists from MITOMAP (v102) with the “blacklisted_site” flag using GATKs VariantFiltration.39 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).43

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 (PAKIYWBMMNF) 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.44 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, antiviral 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 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.49 Next, any variants that were previously found to be likely false positives caused by NuMTs were removed.49,51 For this step we used variants in the RH094 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_freq_log_m < - glm(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: “glm(freq ~ age + (0 + age | patient), data = healthy_freq, family = poisson(link = “log”))”; “glm(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. 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 SE(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 = TRUE, 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, with 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.

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”)), 2))”.

Mutation spectra
Mutation spectra and their cosine similarities were calculated and visualized using the MutationalPatterns (v3.3.4) R package. 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).

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.