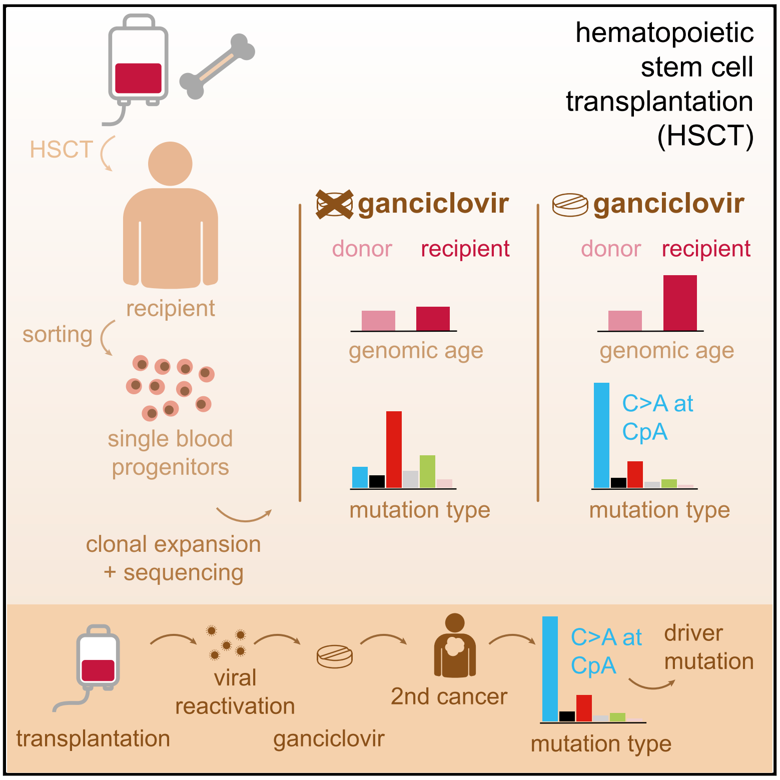
Cell Stem Cell

Clinical and Translational Report

Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients

Graphical abstract



Highlights

- Stem cell transplantation does not enhance mutagenesis in most human recipients
- The antiviral drug ganciclovir causes unique mutation signature in transplanted HSPCs
- Ganciclovir-associated mutagenesis is found in cancers of transplantation recipients
- Genetic drivers in cancers of transplantation recipients can be caused by ganciclovir

Authors

Jurrian K. de Kanter, Flavia Peci, Eline Bertrums, ..., Marc Bierings, Mirjam Belderbos, Ruben van Boxtel

Correspondence

m.e.belderbos@prinsesmaximacentrum.nl (M.B.),
r.vanboxtel@prinsesmaximacentrum.nl (R.v.B.)

In brief

de Kanter et al. demonstrate that antiviral treatment with ganciclovir causes a unique mutational signature in stem cells of human transplant recipients. This signature was also found in therapy-related cancers and can cause cancer driver mutations.

de Kanter et al., 2021, Cell Stem Cell *28*, 1726–1739
October 7, 2021 © 2021 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.stem.2021.07.012

CellPress

CellPress
OPEN ACCESS

Cell Stem Cell

Clinical and Translational Report

Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients

Jurrian K. de Kanter,[1,2,5] Flavia Peci,[1,2,5] Eline Bertrums,[1,2,3] Axel Rosendahl Huber,[1,2] Anaïs van Leeuwen,[1,2] Markus J. van Roosmalen,[1,2] Freek Manders,[1,2] Mark Verheul,[1,2] Rurika Oka,[1,2] Arianne M. Brandsma,[1,2] Marc Bierings,[1,4] Mirjam Belderbos,[1,2,5,*] and Ruben van Boxtel[1,2,5,6,*]

[1]Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, Utrecht 3584 CS, the Netherlands
[2]Oncode Institute, Jaarbeursplein 6, 3521 AL Utrecht, the Netherlands
[3]Department of Pediatric Oncology/Hematology, Erasmus Medical Center, Rotterdam 3015 GD, the Netherlands
[4]Paediatric Blood and Marrow Transplant Program, University Medical Center Utrecht, Utrecht, Netherlands
[5]These authors contributed equally
[6]Lead contact
*Correspondence: m.e.belderbos@prinsesmaximacentrum.nl (M.B.), r.vanboxtel@prinsesmaximacentrum.nl (R.v.B.)

https://doi.org/10.1016/j.stem.2021.07.012

SUMMARY

Genetic instability is a major concern for successful application of stem cells in regenerative medicine. However, the mutational consequences of the most applied stem cell therapy in humans, hematopoietic stem cell transplantation (HSCT), remain unknown. Here we characterized the mutation burden of hematopoietic stem and progenitor cells (HSPCs) of human HSCT recipients and their donors using whole-genome sequencing. We demonstrate that the majority of transplanted HSPCs did not display altered mutation accumulation. However, in some HSCT recipients, we identified multiple HSPCs with an increased mutation burden after transplantation. This increase could be attributed to a unique mutational signature caused by the antiviral drug ganciclovir. Using a machine learning approach, we detected this signature in cancer genomes of individuals who received HSCT or solid organ transplantation earlier in life. Antiviral treatment with nucleoside analogs can cause enhanced mutagenicity in transplant recipients, which may ultimately contribute to therapy-related carcinogenesis.

INTRODUCTION

Life-long production of all mature blood cells is orchestrated by self-renewing, multipotent hematopoietic stem cells (HSCs). Aside from their critical role in homeostatic hematopoiesis, HSCs are the only stem cells that are used routinely for therapeutic purposes. HSC transplantation (HSCT) is performed in more than 40,000 individuals worldwide annually as a curative treatment for bone marrow failure, severe immune deficiency, hemoglobinopathy, inborn errors of metabolism, and leukemia (Pasquini et al., 2010; Passweg et al., 2016). Furthermore, genetically modified HSCs are used increasingly in individuals undergoing gene therapy for monogenic diseases, such as severe combined immunodeficiency, β-thalassemia, and sickle cell anemia, as well as for cancer and HIV/AIDS (Aiuti et al., 2002, 2013; Dunbar et al., 2018; De Ravin et al., 2016; Xu et al., 2019). Because of increased use of HSCT as a treatment strategy as well as improved transplantation protocols, the numbers of HSCT survivors and their life expectancy continue to increase (Bhatia, 2011). Currently, it is estimated that there are more than 500,000 HSCT survivors across the globe, and this number is expected to increase 5-fold by 2030 (Bhatia, 2011; Clark et al., 2016; Majhail et al., 2013). Accordingly, the long-term safety of HSCT, and of stem cell therapy in general, is becoming increasingly important.

A major concern for any clinical therapy using live cells is the presence and acquisition of DNA mutations (Kuijk et al., 2020; Thompson et al., 2020; Yamanaka, 2020). Unwanted mutations may negatively influence the longevity of the administered cell product, alter essential cell functions, or even predispose to malignant transformation. This concern has been particularly related to therapies in which genetically engineered cells or human pluripotent stem cells (hPSCs) are used (Andrews et al., 2017; Avior et al., 2019; Lamm et al., 2016; Thompson et al., 2020; Yamanaka, 2020). For instance, in a clinical trial using autologous induced hPSC-derived retinal cells to treat individuals with macular degeneration, administration of the cell product was abandoned because the cells carried a novel mutation of unknown significance (Mandai et al., 2017). Furthermore, the occurrence of vector-mediated mutagenesis of gene therapy-corrected stem cells has led to international guidelines to maintain the biosafety of this type of therapy and monitor its recipients (Collins and Gottlieb, 2018; Hacein-Bey-Abina et al., 2008; Howe et al., 2008). However, the genomic safety and mutational consequences of the oldest and most frequently applied stem cell therapy, HSCT, remain unknown.

Here we aimed to systematically assess the mutational consequences of HSCT in human recipients, using whole-genome sequencing of individual HSPCs before and after

© 2021 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

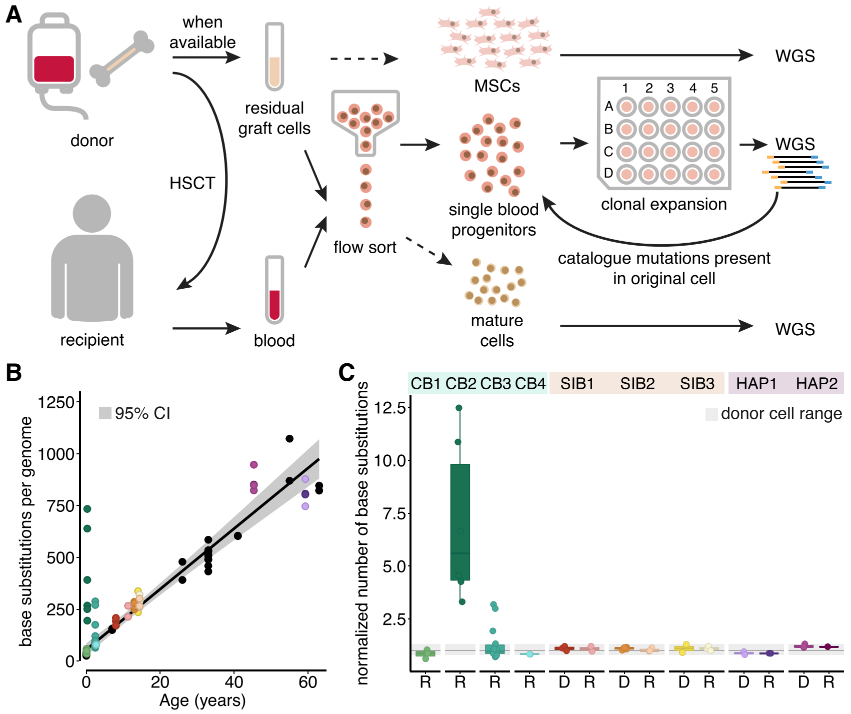


Figure 1. Mutation accumulation associated with HSCT in humans

(A) Schematic of the experimental setup to determine somatic mutations in blood progenitor cells of HSC transplantation (HSCT) donors and recipients.

(B) Correlation between the age and the number of base substitutions per genome in 51 single HSPC clones of 5 HSCT donors and 9 HSCT recipients. Each dot represents a single HSPC clone. A linear mixed effects model of 34 bone marrow clones from 11 healthy individuals (including the HSCT donors) was used to construct the baseline. The 95% CI of the baseline is depicted in gray. HSCT clones are colored similar to (C), and non-HSCT clones of the baseline are shown in black.

(C) The number of base substitutions in donor and recipient HSPC clones shown in (B), normalized to the baseline (expected number of mutations at that age). Each dot represents a single HSPC clone. The range of the normalized number of base substitutions of donor HSPC clones is depicted in light gray. CB, cord blood; SIB, sibling; HAP, haploidentical; D, HSCT donor; R, HSCT recipient.

See also Figure S1 and Tables S1, S2, and S3.

transplantation. For this, we compared the mutation burden in these cells with HSPCs obtained from healthy donors with ages ranging across the entire human lifespan. We demonstrate that the majority of HSCT recipients do not display enhanced mutagenesis. However, multiple HSPCs isolated from two HSCT recipients after transplantation showed an increased mutation burden, which could be attributed to one specific mutational signature. This unique signature is characterized by C > A transversions at CpA dinucleotides with a strong replication strand bias. The same mutational signature was present in six hematologic malignancies, which occurred after HSCT, and in two solid tumors of individuals who underwent renal transplantation earlier in life. These individuals had been treated for viral reactivations after transplantation. By *in vitro* exposure of human umbilical cord blood HSPCs, we prove that this signature is caused by the antiviral nucleoside analog ganciclovir, which is administered to immune-deficient individuals as a first-line treatment of viral reactivation. Our study demonstrates that antiviral treatment with nucleoside analogs after transplantation can be associated with increased mutagenicity, which may ultimately drive development of therapy-related malignancies.

RESULTS AND DISCUSSION

Cataloging somatic mutations in individual HSPCs of human transplantation recipients

We performed whole-genome sequencing (WGS) of clonal HSPC cultures of human HSCT recipients and their donors to catalog all mutations that were present in the parental HSPCs (Figure 1A; Jager et al., 2018; Osorio et al., 2018). We included nine pediatric HSCT recipients who were transplanted with bone marrow cells of an HLA-identical sibling donor (n = 3, sibling 1 [SIB1]–SIB3), a haploidentical parent donor (n = 2, haploidentical 1 [HAP1] and HAP2), or an anonymous umbilical cord blood (UCB) donor (n = 4, cord blood 1 [CB1]–CB4). All recipients had been transplanted for hematologic malignancies after chemotherapy-based myeloablative conditioning. Clinical details are provided in Table S1. We analyzed HSPC clones from residual donor graft cells collected at the time of HSCT and from peripheral blood of the recipient, which was collected 1–295 months after transplantation. At each time point, we analyzed 2–14 HSPC clones per individual by WGS at a depth of 15–30× base coverage. To filter out germline variants, we performed WGS on DNA isolated from donor bone marrow mesenchymal stromal cells (MSCs), bulk T cells, or bulk granulocytes. When a control was unavailable, we used the various clones of the same individual for filtering (STAR Methods; Table S2). The variant allele frequencies (VAFs) of the somatic mutations in all HSPC cultures clustered around 0.5, confirming their clonal origin (Figure S1A). Mutations that accumulated after the first cell division upon plating the single HSPCs will not be shared by all cells in the resulting clonal cultures and were filtered based on their lower VAF (Jager et al., 2018; Osorio et al., 2018; Rosendahl Huber et al., 2019). In total, we identified 15,691 clonal single-base substitutions (SBS) and 927 indels in 51 assessed HSPCs (Tables S2 and S3). We reconstructed phylogenetic trees for all individuals and validated that most mutations in the assessed HSPC clones were acquired independently (Figure S3A). Furthermore, to exclude the possibility that these mutations had been caused by artifacts during library preparation or sequencing, we generated new libraries and re-sequenced the genomes of five clones of two individuals. In total, we could validate 1,049 of 1,070 assessed mutations (overall confirmation rate, 98.0%; range, 96.5%–99.3% per clone; n = 5; Figure S3B).

We detected 365 mutations (2.2% of the total) in coding regions of the genome. None of these were nonsynonymous or truncating mutations in genes that are recurrently mutated in hematological neoplasms. To determine the extent of positive or negative selection that had acted on these clones, we calculated the ratio of non-synonymous to synonymous mutations (dN/dS). The maximum-likelihood estimates of this ratio always included 1, indicating that the HSPCs had undergone neutral selection not only during the *in vitro* culture period but also during life (Figure S1B). We did not observe any acquired structural variations in pre- and post-HSCT clones.

Transplantation-associated mutation accumulation in human HSPCs

We previously established a baseline for mutation accumulation in normal HSPCs across the human lifespan and determined that human HSPCs accumulate about 15 mutations per life year (Osorio et al., 2018). To assess the mutational effect of transplantation, we compared the somatic mutation load in HSPCs collected from human HSCT recipients after transplantation with that of their donor's pre-HSCT clones and with this healthy baseline (Figures 1B, 1C, and S1C). As expected, all pre-HSCT clones fell on the healthy baseline. To compare the post-HSCT clones, we defined the age of these cells as the age of the donor + the interval after HSCT. In the majority of these post-HSCT clones, the number of base substitutions was within the predicted range of normal hematologic aging (ratio observed/expected, 0.6–1.3; Figure 1C). This finding was unexpected because these donor HSPCs have regenerated an entire new blood system in the recipient, which likely requires enhanced proliferation. Nevertheless, these cells did not accumulate additional mutations, apart from those expected to occur because of normal aging. In contrast, in two recipients, we identified 10 independent post-HSCT clones with up to 12-fold more mutations than predicted based on their age (mean observed/expected, 5.15; range, 1.33–12.5; 95% confidence interval [CI], 2.8–7.5; Figures 1B and 1C), which was higher than in any of the pre-HSCT clones. Both HSCT recipients were transplanted with a graft obtained from an UCB donor (Table 1). Consistent with the pediatric age of the subjects in our study, the number of insertions or deletions (indels) was limited and more variable (Figures S1D–S1F). However, the number of indels in single HSPCs was generally within the expected range and did not differ consistently between HSCT donors and their recipients, including post-HSCT clones with a significantly higher base substitution load (Figures S1D–S1F). These data show that, although HSCT is not associated with enhanced mutagenesis in most subjects, there are several HSCT recipients in whom (a subset of) the donor HSPCs accumulate substantial amounts of additional DNA mutations.

Transplantation-associated mutation accumulation can be attributed to a unique mutational signature

Next we aimed to identify the processes underlying HSCT-associated mutagenesis by deciphering mutational signatures from the somatic mutation catalogs of the post-HSCT clones (Figure 2). Such signatures reflect specific mutational processes that have been active during the life of the assessed HSPCs (Alexandrov et al., 2013, 2016; Behjati et al., 2014). In HSPC clones with a normal mutation burden, the spectrum was dominated by C > T transitions, which could be attributed to a previously defined HSPC signature (Figures 2A–2C; Lee-Six et al., 2018; Maura et al., 2019; Osorio et al., 2018). This signature reflects clock-like activity of the predominant mutational process in postnatal HSPCs during healthy life (Hasaart et al., 2020), whose underlying mechanism is still unknown. In contrast, in HSPC clones with an increased number of mutations compared with the normal baseline, C > A transversions were the most abundant mutation type, accounting for 40%–87% of the total number of base substitutions (Figures 2A–2D). The number of C > A transversions in these cells was increased significantly compared with HSPCs with a normal mutation burden (Wilcoxon test, $p < 10{-}e5$; Figure 2E). In fact, the higher the increase in mutation load in these post-HSCT clones, the more their spectra deviate from the mutation spectrum normally observed in healthy HSPCs (Figure 2B), indicative of an underlying mutational process that is not normally active. When considering their trinucleotide context, we noted that the C > A transversions occurred preferentially at CpA dinucleotides (Figures 2D and S2), suggesting a single causative process. Indeed, mutational signature analysis revealed that the increase in mutation load in these recipient HSPCs could be attributed exclusively to a previously unidentified SBS signature, which we called "SBSA" (Figures 2C and 2D; Table S4). SBSA is characterized by C > A transversions (86% of all mutations in SBSA), of which more than 90% are NpC > ApA changes (Figure 2D). SBSA mutations occurred in two of the nine individuals (22%) assessed in this study (CB2 and CB3). Of these, 6 of 6 CB2 clones (100%) and 6 of 14 CB3 clones (43%) harbored SBSA mutations. To establish whether the SBSA mutations in these clones were also propagated to mature blood cell progeny, we sequenced the genomes of bulk-sorted B cells and monocytes of individual CB3. Subsequently, we assessed, for each mutation present in CB3 HSPCs, the VAF in these mature populations. We could detect early mutations (i.e., mutations shared between multiple HSPCs, indicative of an ancestral progenitor) with relatively high VAFs in these bulk populations (Figure 3A). Some of the mutations that were unique to the individual clones could also be detected, albeit at lower VAFs. Interestingly, many of these unique mutations were C > ApA mutations, indicating that SBSA mutations occurred later during life and are propagated to mature progeny (Figure 3B). To confirm that SBSA is distinct from previously defined mutational signatures, we calculated its similarity to the signatures from the catalogue of somatic mutations in cancer (COSMIC) database (v.3.0) as well as with *in vitro* established signatures of environmental agents (Kucab et al., 2019; Tate et al., 2019). A cosine similarity of 0.95 or more was used to indicate that two patterns are similar (Blokzijl et al., 2018). We found that SBSA did not match any of the previously defined mutation signatures (Figure 2F). SBSA showed highest cosine similarity with, but was still distinct from, SBS38, SBS18, and a potassium bromate ($KBrO_3$)-induced signature (cosine similarity of 0.83, 0.57, and 0.81, respectively; Figures 2F, 4A, and S4C).

Molecular characterization of SBSA

SBS38, SBS18, and the $KBrO_3$ signature have been attributed to oxidative stress-induced mutagenesis, which is thought to be

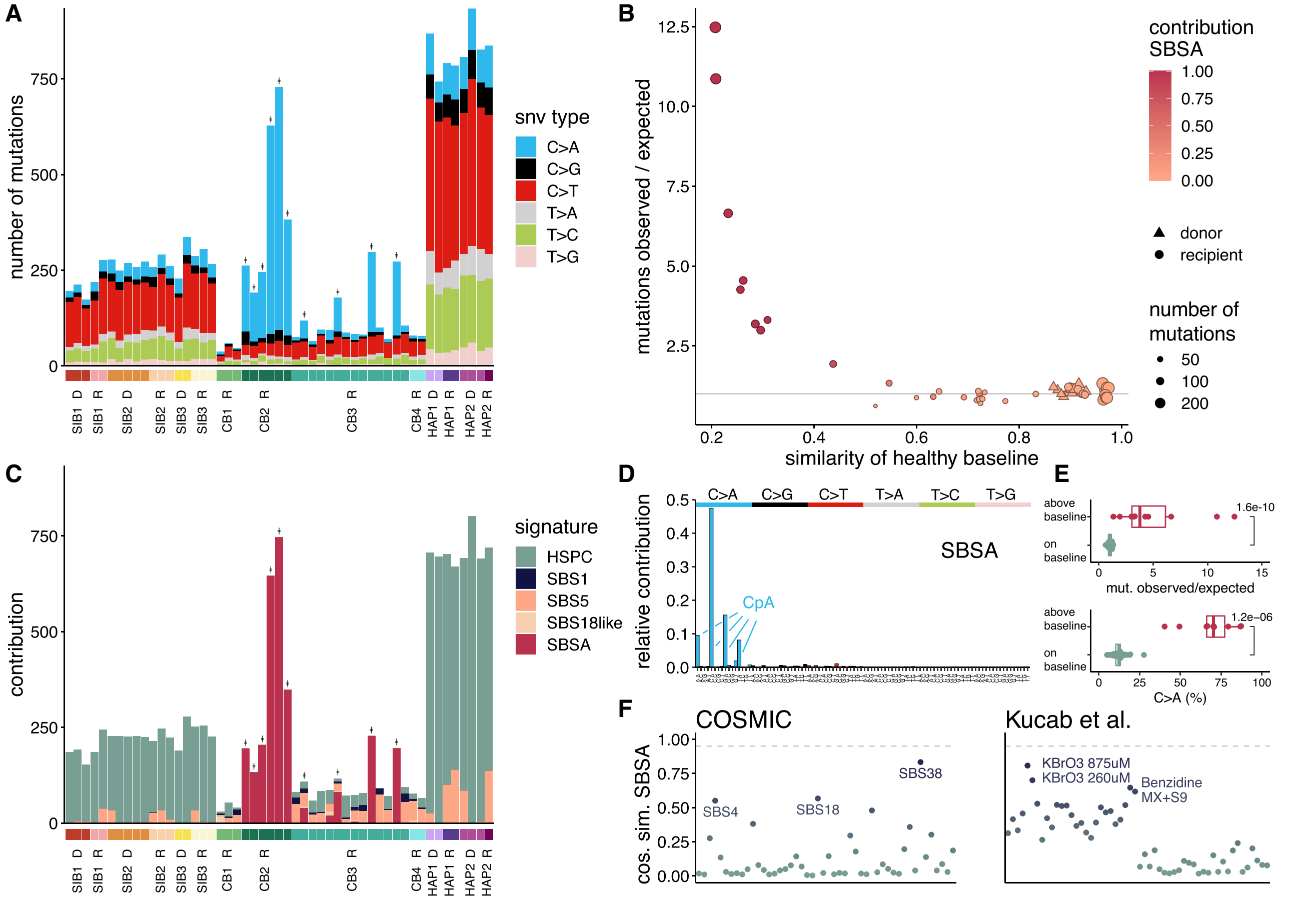


Figure 2. Transplantation-associated mutagenesis can be attributed to a unique mutational signature, SBSA

(A) Single base substitution (SBS) mutational spectra from HSCT donor and recipient HSPCs. “” indicates recipient HSPCs with an increased mutational burden. For the 96-trinucleotide mutational profiles of the individual cells, see Figure S2.

(B) Age-adjusted number of mutations in each single HSPC clone (dot/triangle) compared with its similarity to the healthy baseline. Similarity was calculated as the cosine similarity of the 96-trinucleotide profiles. The colors of the symbols indicate the contribution of SBSA to the mutational profile of the HSPCs in the refitting analysis depicted in (C).

(C) The contribution of the five signatures found by non-negative matrix factorization (NMF) to the mutational profile of each HSPC.

(D) SBS 96-trinucleotide mutational signature of SBSA as inferred by NMF of the HSCT donor and recipient HSPCs. See also Table S4.

(E) The ratio of observed versus expected mutations of HSCT HSPC clones with SBSA mutations that have an increased mutation load and of HSCT HSPC clones that lie on the age line (top, Wilcoxon test) and the percentage of mutations that are a C > A transversion of the same groups of clones (bottom, Wilcoxon test).

(F) The cosine similarity between the SBSA signature and SBS mutational signatures from the COSMIC v.3.0 database and *in vitro* established signatures of environmental agents (Kucab et al., 2019).

driven by 8-oxo-guanine lesions in the DNA and subsequent mispairing of this damaged base with adenine during replication (Alexandrov et al., 2013; Brem et al., 2017; Kucab et al., 2019). To determine whether SBSA also reflects oxidative stress-induced mutagenesis, we compared several genomic characteristics of these mutational signatures. First, because some known mutational processes preferentially target a DNA context broader than 3 bases (Pleguezuelos-Manzano et al., 2020), we assessed the 10 bases up- and downstream of the C > ApA mutations of SBSA. We compared this context with oxidative stress-induced C > A transversions caused by $KBrO_3$ (Kucab et al., 2019) and a knockout of *OGG1* (OGG1KO), which has a central role in 8-oxo-guanine base excision repair (Boiteux et al., 2017; Figures 4A and S4). C > ApA mutations in HSPCs with SBSA were consistently associated with an increased incidence of cytosines at position −1 and −6, of guanines at position −2, and of thymines at position −3 (Figures 4B and S4A). In contrast, this context did not occur in the $KBrO_3$ and OGG1KO C > ApA mutations, suggesting a different mutagenic cause of SBSA.

In post-HSCT clones with high mutation load and contribution of SBSA, C > A transversions demonstrated a highly significant Watson-versus-Crick-strand lesion segregation (false discovery rate [FDR] < 10e−12), which was absent in cells treated with $KBrO_3$, deficient for *OGG1*, and in HSPCs with a normal baseline mutation load (FDR = 0.17, 0.29, and 0.48, respectively; Figures 4C, 4D, and S4E). It has been shown previously that such lesion segregation reflects accumulation of mutagenic DNA lesions within a single cell cycle, which causes strand-specific segregation of these lesions into daughter cells (Aitken et al.,

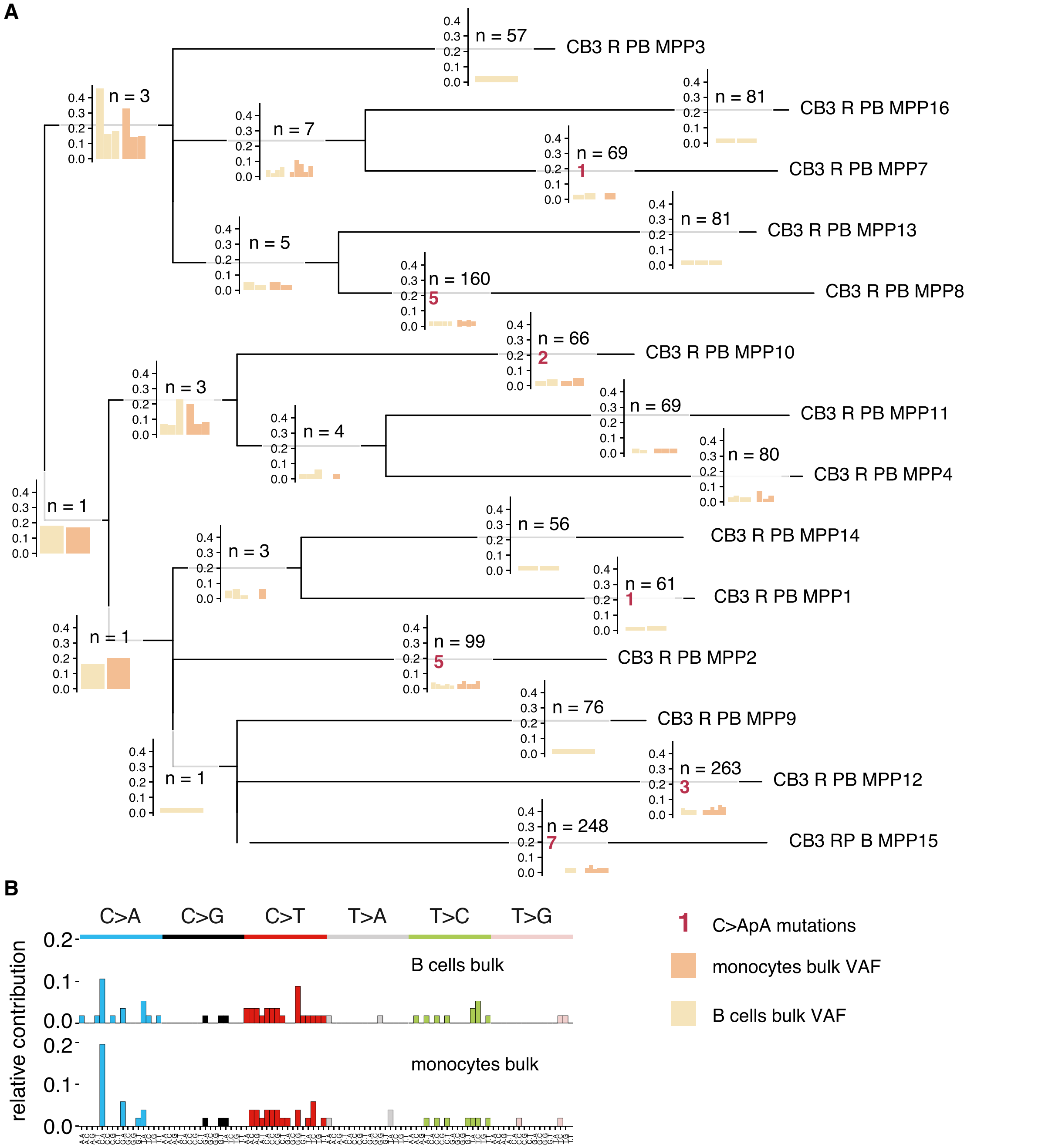


Figure 3. Detection of HSPC mutations in bulk mature populations

(A) The phylogenetic tree of the HSPCs of individual CB3. At each branch, a bar graph is plotted. The number above each bar graph indicates the total number of mutations in that branch. Each bar represents the VAF of a mutation in that branch of the tree in WGS data of the bulk-sorted B cells or monocytes of CB3. Each bar represents a single mutation that is found in that mature population. Mutations that are not found in the mature populations are not shown.

(B) The 96-trinucleotide profile of all HSPC mutations that are found in each of the mature populations.

For the phylogenetic trees of all individuals, see Figure S3.

2020). As a result, one daughter cell and its progeny only carry mutations on the Watson or the Crick strand, whereas the other daughter cell and its progeny carry mutations in the other strand. These data suggest that the causative process of SBSA operates during a short period of time, possibly even a single cell division.

Next we assessed whether SBSA mutations are associated with DNA transcription or replication. SBSA mutations showed a small bias toward the transcribed strand (FDR = 0.016), but they did not show enrichment in exons or gene bodies (FDR = 0.11), suggesting that transcription-coupled repair can resolve the DNA lesions causing SBSA but is likely not the main repair mechanism (Figures S4B and S4F; Haradhvala et al., 2016; Tomkova et al., 2018). SBSA mutations were slightly depleted in late-replicating regions of the DNA (FDR < 10e−4; Figure 4E), suggesting that the mutagenic cause or involved repair process is not strongly linked to replication timing. We noted that SBSA C > A transversions showed a significant replication strand bias toward the leading strand (FDR < 10e−23; Figures 4F and S4D), which indicates that the mutagenic process underlying SBSA is directly coupled to DNA replication (Haradhvala et al., 2016; Tomkova et al., 2018). These data suggest that, unlike oxidative-stress induced mutations, SBSA mutations in post-HSCT clones are caused by erroneous DNA replication upon short-term exposure to a mutagenic source.

SBSA is caused by the antiviral nucleoside analog ganciclovir

To identify the mutagenic source of SBSA, we analyzed the clinical data of our transplant recipients (Table S1). Both HSCT recipients who harbored SBSA-positive HSPCs (CB2 and CB3) had developed early viral reactivations after transplantation, which required treatment with the antiviral drugs foscarnet (FC) and (val)ganciclovir (GCV) (Table S1). Interestingly, GCV is a synthetic analog of 2′-deoxy-guanine and a competitive inhibitor of deoxyguanosine triphosphate (dGTP) incorporation into DNA (Seley-Radtke and Yates, 2018). FC is a pyrophosphate analog thought to directly inhibit viral polymerase activity (Crumpacker, 1992). Because these compounds affect DNA replication, they are likely candidates for causing SBSA mutations. To test this, we exposed human $CD34^+$ UCB HSPCs to GCV and/or FC *in vitro* (Figure 5A). Although GCV caused dose-dependent cell death at micromolar concentrations, which are also observed in human plasma (IC_{50}, 4.64 μM; Piketty et al., 2000), FC did not induce cell death at any of the tested concentrations (Figure 5B). We then treated these cells for 24 h with 5 μM GCV and/or, similar to previous publications, a 40 times higher concentration of FC (200 μM; Maggs and Clarke, 2004). GCV and the combination treatment caused substantial DNA damage, visualized by γ-H2AX staining, whereas FC exposure alone did not cause considerable cell death (Figures 5C, 5D, and S5C). To assess the mutational consequences caused by these antiviral drugs, we subsequently performed a clonal expansion step and performed WGS on 2–3 clones for each condition. HSPCs exposed to GCV or to the combination therapy showed increased numbers of SBSs compared with HSPCs exposed to FC alone or untreated clones, with a bias toward C > A transversions (Figure 5E). The number of indels was similar between GCV-, FC-, and control-treated cells; no copy number variations or structural rearrangements were found (Figures S5A and S5B). Importantly, the 96-trinucleotide profile induced by *in vitro* exposure to GCV was essentially identical to SBSA found in affected individuals (cosine similarity, 0.999; Figure 5F). Similar to SBSA, the C > A mutations induced by *in vitro* GCV exposure (and by GCV+FC) were strongly biased toward the leading replication strand as well as the transcribed strand, were depleted in late-replicating regions, showed strong lesion strand segregation, and had a similar extended base context as SBSA (Figures S5D–S5H). These data clearly demonstrate that GCV is the cause of the SBSA mutations.

SBSA mutations in cancer

Accumulation of somatic mutations is a key mechanism promoting carcinogenesis. To assess whether SBSA mutations can contribute to cancer development, we determined its presence in the genomes of allogeneic and autologous HSCT donors and recipients (Boettcher et al., 2020; Gondek et al., 2016; Husby et al., 2020; Lombard et al., 2005; Mouhieddine et al., 2020; Ortmann et al., 2019; Figure 6). To enable detection of SBSA in these datasets, we developed a random forest (RF) classifier. This machine learning technique employs the previously defined features of SBSA to predict whether a SBS originates from SBSA (Figures S6A, S6B, and S6G). We trained the RF on pre- and post-HSCT HSPCs and on the healthy baseline HSPCs depicted in Figure 1. Importantly, the RF classifier assigned the highest importance to nucleotides that were present on the +1, −1, and −2 positions surrounding the C > A-mutated cytosine, underlining the importance of the broader sequence context of SBSA mutations. To prevent false-positive calls, we applied the RF to 1,000 sets of randomly generated base substitutions. The highest percentage of SBSA-positive mutations in these random datasets was used to select the cutoff for ‘‘true’’ SBSA positivity, which was 2.3% (Figure S6G). To validate the resulting RF and the applied cutoff, we tested its performance on a control WGS dataset of HSPCs of a 60-year-old healthy individual (Lee-Six et al., 2018) and on a dataset of clonal hematopoiesis of indeterminate potential (CHIP) mutations in bulk WGS of 97,691 healthy individuals (Bick et al., 2020; Figure 6C). As expected, the RF identified less than 1% of SBSA-positive mutations in both datasets, confirming the specificity of this classifier.

Next we applied this RF classifier to sequencing datasets of human metastatic cancers (n = 3,668) (Priestley et al., 2019) and of hematologic disorders after allogeneic and autologous HSCT, such as clonal hematopoiesis (n = 290) (Boettcher et al., 2020; Husby et al., 2020; Mouhieddine et al., 2020; Ortmann et al., 2019), therapy-related neoplasms (n = 9) (Berger et al., 2018; Gondek et al., 2016), and relapsed acute myeloid leukemia (AML) after allogeneic HSCT or chemotherapy (n = 44) (Christopher et al., 2018; Stratmann et al., 2021). In total, the RF classified nine cancers of nine individuals as SBSA positive (Figure 6; Table 1). The first was therapy-related AML (tAML; PMC11396), in which SBSA had an estimated contribution of 28% (Figures 6A and 6B). This individual had received allogeneic HSCT for relapsed acute lymphoblastic leukemia (ALL) with successful engraftment but developed tAML of patient origin 3 years later (Table 1). Using the RF classifier on WGS data of this individual's tAML, the primary ALL, as well as three normal HSPCs collected 3 months prior to HSCT, we found that only the tAML was classified as SBSA positive (Figure 6A). This finding was confirmed using mutational signature analysis (Figure 6B), the ±10 nt context (Figure S6C), and replication strand bias (Figure S6H). The C > A mutations did, however, not display a Watson-versus-Crick bias (Figure S6K). Although five mutations were shared between the tAML and one of the healthy HSPCs collected prior to transplantation (Figure S6F), none of these were C > ApA mutations. In line with our *in vitro* findings, the individual was treated with FC and GCV for cytomegalovirus (CMV) reactivation after HSCT.

The second SBSA-positive tumor was a donor cell leukemia (DCL), reported in a study by Gondek et al. (2016) on clonal

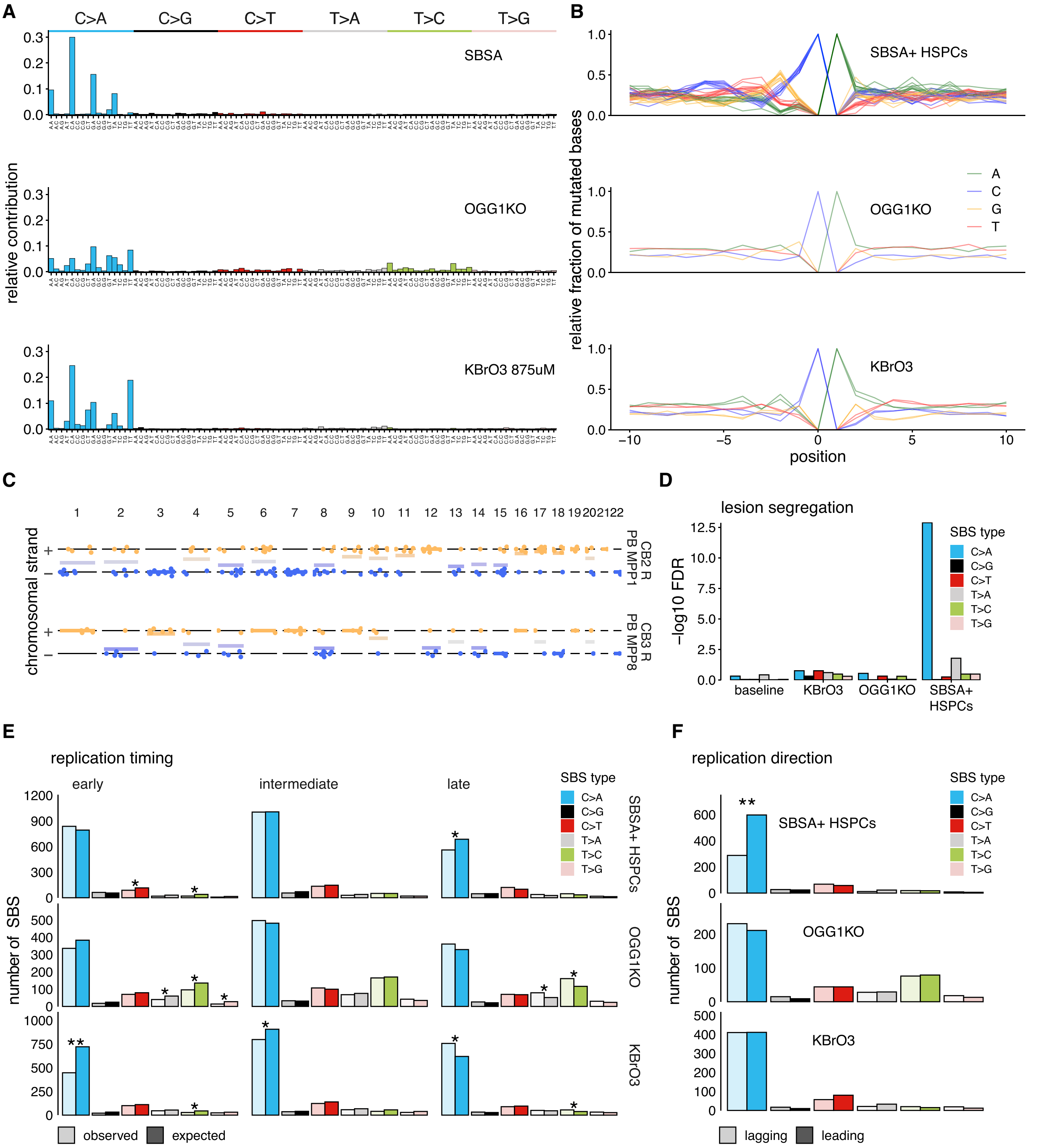


Figure 4. SBSA is characterized by lesion segregation and a strong replication direction bias

(A) SBS 96-trinucleotide mutational profiles of SBSA and oxidative stress-associated signatures of exposure to $KBrO_3$ or knockout of *OGG1*.

(B) The −10:+10 nucleotide context of C > ApA mutations of five SBSA-positive HSPC clones, knockout of *OGG1*, and two $KBrO_3$-treated clones. Each line represents the mutation context in a single clone. Position 0 and 1 contain the C > A and subsequent A of the C > ApA mutations, respectively.

(C) The chromosomal strand and position of the cytosine of C > A mutations of two clones positive for SBSA.

(D) FDR-corrected p values of Wald-Wolfowitz runs tests on summed numbers of mutations and runs in each group.

(E) Enrichment/depletion of SBSA-positive HSPC clones, knockout of *OGG1*, and exposure to $KBrO_3$ in early-, intermediate-, and late-replicating regions. *FDR < 0.05. **FDR < 10^{-7}

(F) Replication strand bias of the same data as depicted in (E).

See also Figure S4.

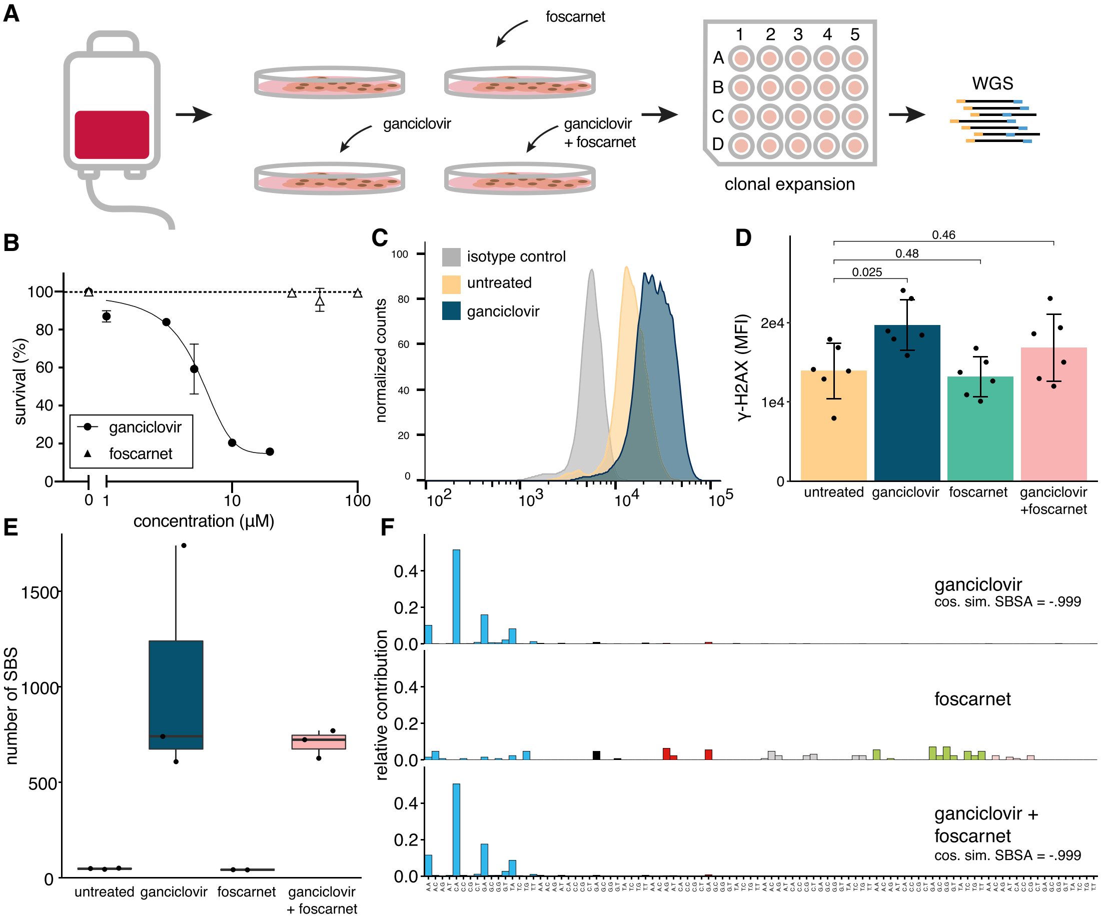


Figure 5. Ganciclovir induces SBSA mutations *in vitro*

(A) Experimental setup of *in vitro* treatment of CD34+ human UCB cells with the antiviral agents foscarnet [FC], ganciclovir [GCV], and a combination of both. After 24 h of treatment, single clones are sorted into 96-well plates, expanded, and whole genome sequenced.

(B) Survival curve and ganciclovir treatment. For FC, no curve could be fitted because of the low percentage of cell death. 200 μM FC is not shown and caused 86% survival.

(C) Representative histogram of γ-H2AX intensity of isotype, untreated, and ganciclovir-treated CB cells.

(D) The γ-H2AX mean fluorescence intensity (MFI) of three CB samples, each treated with each condition twice (Wilcoxon test). See Figure S6C for values per sample and a positive radiation control.

(E) The number of SBSs of each of the treatment conditions (5 μM ganciclovir and/or 200 μM FC).

(F) 96-tri-nucleotide profiles of each treatment condition. The mutations of the untreated condition are subtracted from each profile to normalize for *in-vitro*-acquired mutations.

See also Figure S5.

hematopoiesis after HSCT and its progression toward malignancy (Figures 6C and 6D). This individual (hereafter called Gondek1) was transplanted for AML and developed a DCL 3.5 years after HSCT. The mutation profile of this DCL scored high in the RF (Figure 6C) and had a clear SBSA signature (Figure 6D), whereas the graft material of this individual, collected before HSCT, did not have these mutations.

Moreover, 4 of 44 assessed AML relapses were SBSA positive, and all individuals had been transplanted (Figures 5C and 5D; Christopher et al., 2018). Again, we confirmed this with mutational signature analysis, replication direction bias, and the extended context (Figures S6E and S6J). Also, in this case, the C > A mutations did not have a Watson-versus-Crick asymmetry (Figure S6K). For 3 of 4 individuals, the medical history could be obtained (Table 1). All three individuals developed early CMV reactivation after HSCT and received GCV as antiviral treatment, consistent with an approximate prevalence of SBSA in 14% (4 of 29 relapses after HSCT; 95% CI, 4%–29%) of AML relapses after allogeneic HSCT.

Finally, the RF classified three tumors from a Dutch collection of 3,668 solid cancer metastases as SBSA positive (Priestley et al., 2019; Figure 6E). All three were liver metastases of solid tumors (melanoma, breast carcinoma, and vulva carcinoma). Intriguingly, although none of these individuals had received an HSCT, two of three individuals had received a kidney transplantation earlier in life. For one of these individuals, we could retrieve the treatment history, which revealed that the individual received GCV to treat a viral reactivation after the transplantation. Further

Table 1. Clinical information for SBSA-positive cancers

Sample	Primary diagnosis	Transplantation	(Second) cancer	Viral reactivations	Antiviral therapy	(Second) cancer driver mutations (C > ApA)	Reference
11396 – Dx2 AML	ALL	HSCT	AML	CMV	GCV FC		N/A
633734 – relapse	AML	HSCT	AML relapse	CMV	GCV	*NRAS* p.Q61K	Christopher et al., 2018
103342 – relapse	AML	HSCT	AML relapse	CMV	GCV, valganciclovir		Christopher et al., 2018
814916 – relapse	AML	HSCT	AML relapse	CMV	ganciclovir		Christopher et al., 2018
AML_015	AML	HSCT	AML relapse	unknown	unknown		Stratmann et al., 2021
Gondek1 – DCL	AML	HSCT	DCL	unknown	unknown	*SETBP1* p.T873K	Gondek et al., 2016
CPCT02090030T	renal insufficiency	kidney Tx	vulvar carcinoma metastasis	unknown	unknown	*HRAS*, p.Q61K	Priestley et al., 2019
CPCT02110076T	renal insufficiency	kidney Tx	breast carcinoma metastasis	CMV	valganciclovir		Priestley et al., 2019
CPCT02340067T	melanoma	none	melanoma relapse metastasis	none	none		Priestley et al., 2019

ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; HSCT, hematopoietic stem cell transplantation; Tx, transplantation; CMV, cytomegalovirus; FC, foscarnet; GCV, ganciclovir.

analyses confirmed the SBSA ± 10 nt context and replication strand bias in the metastases of these two transplanted individuals but showed no Watson-versus-Crick asymmetry (Figures 6F, S6D, S6I, S6K, and S6L). In contrast, the tumor of the non-transplanted individual with melanoma did not show this context nor bias (Figures S6D and S6I) and is therefore considered a false positive result of the RF.

Three of the driver mutations in the SBSA-positive tumors (*SETBP1* T873K in Gondek1, *HRAS* Q61K in CPCT02090030T, and *NRAS* Q61K in 633734) were C > ApA transversions, suggesting a direct contribution of SBSA to cancer development in these individuals. We estimated the probability of SBSA having caused these mutations using a method published previously (Morganella et al., 2016). The three mutations had a probability of 84% (*SETBP1*), 90% (*HRAS*), and 97% (*NRAS*) to be caused by SBSA. To test the overall damage potential of SBSA, we calculated the enrichment of stop-gain, missense, and synonymous mutations SBSA can potentially cause in 38 blood cancer driver genes in the human genome to a background of random mutations and compared this with SBS18, $KBrO_3$, and clock-like mutational signatures (Figure 6G). These calculations showed an increased potential of SBSA to cause stop-gain mutations (ratio of 1.6 for stop-gain compared with background) (Figure 6G). However, this analysis does not take into account DNA accessibility, DNA folding, and other extrinsic factors. To address this issue, we calculated what percentage of hematologic cancer driver mutations in the COSMIC dataset could arise because of SBSA (Jaiswal et al., 2014). Of these hematologic cancer drivers, 7.8% of stop-gain mutations were caused by C > ApA mutations, whereas only 2.8% of non-synonymous mutations occurred in the SBS context, confirming our previous results (Figure 6H). These results identify the presence of the GCV-induced mutational signature in several types of cancers of human transplantation recipients and demonstrate its potential to cause cancer driver mutations; in particular, stop-gain mutations.

In this study, we provide insight into the effect of HSCT on the acquisition and causative processes of somatic mutations in the transplanted stem cells, and into their effect on malignant transformation. During normal human aging, HSCs are estimated to acquire 14–15 SNVs per year (Hasaart et al., 2020; Lee-Six et al., 2018). Because HSCs divide approximately every 40 weeks (Catlin et al., 2011), this would mean that, if all mutations occur because of stochastic replication errors, then each HSC acquires 11 mutations per division. If 1,000–5,000 transplanted HSCs would repopulate the new blood system and regenerate the estimated average pool of 200,000 HSCs, then this would mean they each need to divide 5–8 times (Lee-Six et al., 2018). This would result in ~60–80 more mutations per cell. However, the majority of transplanted HSPCs in our study did not display an enhanced mutation burden. There may be several reasons for this finding. Post-transplantation hematopoietic reconstitution is likely mediated by distinct HSPC subsets, perhaps reducing the proliferative demand on the most primitive HSPCs (Biasco et al., 2016; Scala et al., 2018). Furthermore, current estimates of the human HSPC pool are based on steady-state hematopoiesis, whereas the number of HSPCs that contribute to blood formation (and the number of cell divisions needed to regenerate the system) may differ between homeostatic hematopoiesis and hematopoietic regeneration (Lu et al., 2019; Sun et al., 2014; Weissman, 2000). Finally, as suggested in recent studies, the number of mutations that accumulate in HSPCs as a result of errors during cell division may be quite low, and time is likely to be the most important determinant of mutation load (Abascal et al., 2021; Lee-Six et al., 2018; Osorio et al., 2018).

Importantly, although we did not observe a general mutational increase in all HSCT recipients, we do show that treatment of

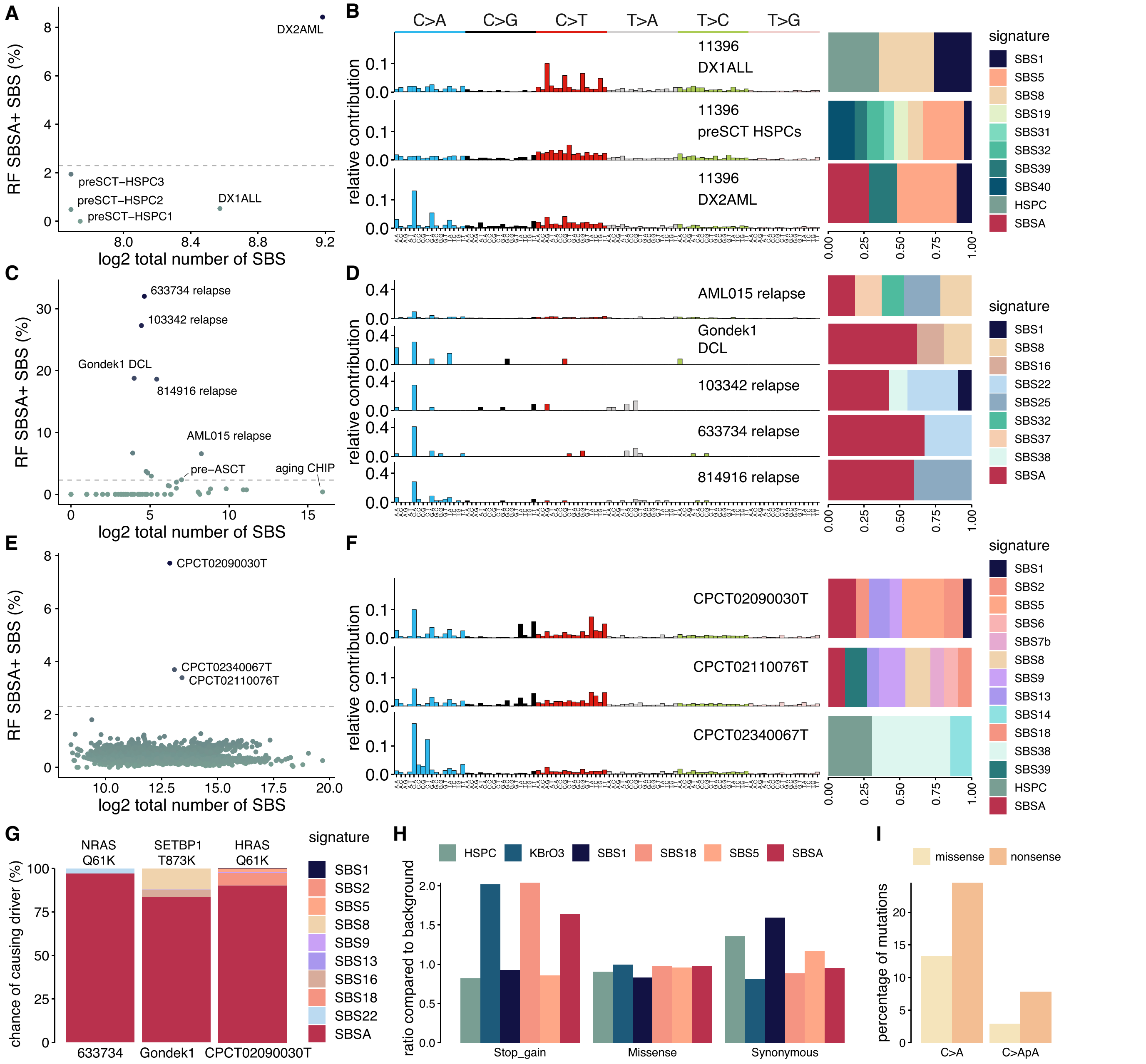


Figure 6. SBSA is present in transplant-related cancers and can cause cancer driver mutations

(A, C, and E) The percentage of RF-predicted SBSA mutations compared with the total number of mutations in samples of (A) individual PMC11396; (C) targeted and WGS mutation datasets of autologous and allogeneic HSCT grafts and recipients, normal aging, age-associated CHIP, post-HSCT AML relapses, and post-HSCT tMN cases; and (E) a Dutch WGS cohort of 3,668 solid tumor metastases (Priestley et al., 2019). In (C), only samples with more than 1 positive mutation are labeled.

(B) The SBS 96-trinucleotide mutational profiles of the primary ALL, pre-SCT HSPC clones (pulled), and tAML of individual PMC11396.

(D) Similar to (B) but of the SBSA-positive samples from (C) (Gondek et al., 2016). DCL, donor cell leukemia.

(F) Similar to (B) but of metastases that are SBSA-positive, predicted by the RF in a Dutch cohort of 3,668 solid tumor metastases from (E) (Priestley et al., 2019).

(G) Probability estimation of each signature in a tumor causing C > ApA driver mutations.

(H) The potential mutational effect of six SBS mutational signatures, including SBSA, in blood cancer driver genes, normalized to a "flat" background signature with equal contribution of all SBS 96-trinucleotide mutation types.

(I) The percentage of COSMIC cancer driver SBS mutations in blood cancer driver genes that are C > A mutations or C > ApA mutations.

See also Figure S6 and Table 1.

post-transplantation viral reactivations with GCV causes a substantial increase in the mutational burden and a unique SBS signature in the transplanted HSPCs. We also identified SBSA in six hematologic malignancies that developed after HSCT as well as in two solid tumor metastases of individuals who had received a kidney transplant previously, supporting the concept that GCV-associated mutagenesis may contribute to development of malignancies after transplantation (hematological or solid). Indeed, we identified 3 driver mutations in these malignancies, which could be attributed to SBSA with a high likelihood. In general, mutations attributed to SBSA have a similar chance of being missense mutations compared with age-related signatures (i.e., SBS1, SBS5, and the HSPC signature), but a 1.6 times higher chance of being a nonsense mutation. In contrast, we observed neutral drift for nonsense mutations in SBSA-positive HSPCs. Therefore, the enhanced rate of nonsense mutations by ganciclovir-induced mutagenesis was at a rate below our detection limit and did not lead to strong positive selection. GCV is a 2′-deoxy-guanine analog that competes with dGTP for DNA incorporation, after which it is thought to inhibit DNA replication (Chen et al., 2014). However, antiviral nucleoside analogs have also been reported to mediate their effect by inducing lethal mutagenesis of the viral genome (Loeb et al., 1999). Importantly, our data show that GCV is also highly mutagenic to the human host DNA and provide insight into how GCV induces mutations in human cells. GCV predominantly causes C > A changes at CpA dinucleotides. The transcriptional strand bias of GCV-induced mutations would be in line with a guanine adduct-blocking transcription. Because GCV is a guanine analog, one of the potential explanations would be that SBSA mutations are caused by incorporation of the antiviral compound into the DNA during replication. This would be a possible explanation for why only part of the HSPCs of CB3 harbor SBSA mutations. Following this hypothesis, if some HSPCs were cycling during GCV exposure and others were not, only the former would accumulate more SBSA mutations. Because the SBSA mutations in the transplanted HSPCs displayed a Watson-versus-Crick bias, the underlying lesions are not always resolved within one replication cycle, in line with the idea that GCV is incorporated in the DNA. We did not observe the Watson-versus-Crick strand asymmetry in the SBSA-positive tumor samples, which generally had a higher number of mutations attributed to signatures other than SBSA. This highlights the usefulness of studying children, in whom the number of background mutations is low and any SBS signature thus more pronounced. Finally, the replication strand asymmetry indicates that, if GCV would be incorporated, then this would occur more efficiently during lagging DNA strand synthesis (Tomkova et al., 2018). However, our data are not definitive proof of this mechanism underlying GCV-induced mutagenesis and repair of GCV-induced lesions.

GCV is used for prevention and first-line treatment of CMV disease in transplantation recipients as well as in individuals with congenital CMV infection and CMV reactivation in those with severe immune deficiency or HIV/AIDS (Griffiths and Lumley, 2014). Therefore, its mutational consequences are likely to have a more widespread healthcare effect than only in transplantation recipients. The mutagenic effect of GCV and its long-term clinical consequences should be assessed in large cohorts. Furthermore, we demonstrate that GCV-induced mutations are not only observed in human HSPCs and leukemia but also in solid tumors of different tissue origins, indicating that GCV can be mutagenic for multiple cell types in the human body. Consequently, GCV-induced mutagenesis in other tissues needs to be investigated to fully characterize the contribution of this antiviral nucleoside analog to carcinogenesis.

In conclusion, our study demonstrates that treatment of human transplantation recipients with the antiviral compound GCV can lead to increased mutation accumulation, which may ultimately contribute to carcinogenesis. In contrast, FC, which is often used interchangeably with GCV, is not mutagenic, potentially providing a safer alternative. Our study emphasizes the clinical relevance of stem cell therapy-associated mutagenesis in humans and urges careful surveillance of HSCT recipients to detect and prevent long-term morbidity.

Limitations of the study

First, although use of *in vitro* clonal expansion allows us to catalog genome-wide mutations in single HSPCs, it may preferentially select HSPCs with enhanced proliferative capacity. We show that the assessed clones had undergone neutral selection for missense and nonsense mutations. In addition, we show that HSPCs with GCV-induced DNA damage still grow out *in vitro*, allowing their detection in our assay. However, we cannot exclude the possibility that other kinds of damage might alter clonal outgrowth efficiency and therefore influence which clones are sequenced.

Second, given that a healthy individual has about 200,000 HSPCs (Lee-Six et al., 2018), the number of HSPCs sequenced for each subject is limited. Although the vast majority of HSPCs in non-GCV-treated HSCT recipients had a normal mutation load, it cannot be excluded that 1 or a few non-assessed HSPCs did acquire additional HSCT-related mutations.

Finally, we show that GCV, a drug that is frequently administered after HSCT, can be mutagenic. Additional research is required to pinpoint the precise mechanism underlying GCV mutagenesis and the repair of GCV-induced lesions. Also, the mutagenic effect of GCV and its long-term clinical consequences should be assessed in large cohorts. Similarly, induced mutagenesis in other tissues needs to be investigated to fully characterize the contribution of this antiviral nucleoside analog to carcinogenesis. Because HSCT is a heterogeneous procedure with many genotoxic exposures, we cannot exclude the possibility that other transplantation-related events that are not covered in our cohort may induce mutations in a subgroup of HSCT recipients.

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
  - ○ HSCT donor/recipient bone marrow and blood
- METHOD DETAILS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.stem.2021.07.012.

ACKNOWLEDGMENTS

The authors would like to thank the Hartwig Medical Foundation (Amsterdam, the Netherlands) for facilitating low-input whole-genome sequencing scripts. We thank Prof. Holmfeldt, Prof. DiPersio, Dr. Christopher, and Dr. Lolkema for sharing additional clinical data and Jan J. Molenaar and M.L. van den Boogaard for providing the knockout data of *OGG1*. We are grateful for the insightful counsel of Prof. Dr. Hans Clevers. We thank the Central Biobank of the UMC Utrecht for providing umbilical cord blood samples. Finally, we thank all HSCT recipients and their donors for participation in this study. This study was financially supported by a VIDI grant of the Netherlands Organization for Scientific Research (NWO) (016.Vidi.171.023 to R.v.B.), a consolidator grant from the European Research Council (ERC) (864499 to R.v.B.), a John Hansen research grant from the DKMS, and a European Society for Blood and Marrow Transplantation Leukemia Fellowship Grant (to M.E.B.).

AUTHOR CONTRIBUTIONS

Conceptualization, M.E.B. and R.v.B.; methodology, M.E.B. and R.v.B.; software, J.K.d.K., F.M., M.J.v.R., R.O., and R.v.B.; formal analysis, J.K.d.K., M.E.B., M.J.v.R., R.O., and R.v.B.; investigation, A.M.B., A.R.H., E.B., M.E.B., F.P., and A.v.L; writing – original draft, M.E.B., J.K.d.K., and R.v.B.; supervision, M.B. and R.v.B; funding acquisition, M.E.B. and R.v.B.

DECLARATION OF INTERESTS

A.R.H., A.v.L., and R.v.B. are named as inventors on a patent application filed resulting from this work.

Received: January 11, 2021
Revised: June 11, 2021
Accepted: July 29, 2021
Published: September 7, 2021

REFERENCES

Abascal, F., Harvey, L.M.R., Mitchell, E., Lawson, A.R.J., Lensing, S.V., Ellis, P., Russell, A.J.C., Alcantara, R.E., Baez-Ortega, A., Wang, Y., et al. (2021). Somatic mutation landscapes at single-molecule resolution. Nature *593*, 405–410.

Aitken, S.J., Anderson, C.J., Connor, F., Pich, O., Sundaram, V., Feig, C., Rayner, T.F., Lukk, M., Aitken, S., Luft, J., et al.; Liver Cancer Evolution Consortium (2020). Pervasive lesion segregation shapes cancer genome evolution. Nature *583*, 265–270.

Aiuti, A., Slavin, S., Aker, M., Ficara, F., Deola, S., Mortellaro, A., Morecki, S., Andolfi, G., Tabucchi, A., Carlucci, F., et al. (2002). Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. Science *296*, 2410–2413.

Aiuti, A., Biasco, L., Scaramuzza, S., Ferrua, F., Cicalese, M.P., Baricordi, C., Dionisio, F., Calabria, A., Giannelli, S., Castiello, M.C., et al. (2013). Lentiviral hematopoietic stem cell gene therapy in patients with Wiskott-Aldrich syndrome. Science *341*, 1233151.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. Science *354*, 618–622.

Andrews, P.W., Ben-David, U., Benvenisty, N., Coffey, P., Eggan, K., Knowles, B.B., Nagy, A., Pera, M., Reubinoff, B., Rugg-Gunn, P.J., and Stacey, G.N. (2017). Assessing the Safety of Human Pluripotent Stem Cells and Their Derivatives for Clinical Applications. Stem Cell Reports *9*, 1–4.

Avior, Y., Eggan, K., and Benvenisty, N. (2019). Retraction. Cell Stem Cell *28*, 173.

Bates, D., Mächler, M., Bolker, B.M., and Walker, S.C. (2015). Fitting linear mixed-effects models using lme4. J. Stat. Softw. *67*, 1–48.

Behjati, S., Huch, M., van Boxtel, R., Karthaus, W., Wedge, D.C., Tamuri, A.U., Martincorena, I., Petljak, M., Alexandrov, L.B., Gundem, G., et al. (2014). Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature *513*, 422–425.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. *57*, 289–300.

Berger, G., Kroeze, L.I., Koorenhof-Scheele, T.N., de Graaf, A.O., Yoshida, K., Ueno, H., Shiraishi, Y., Miyano, S., van den Berg, E., Schepers, H., et al. (2018). Early detection and evolution of preleukemic clones in therapy-related myeloid neoplasms following autologous SCT. Blood *131*, 1846–1857.

Bhatia, S. (2011). Long-term health impacts of hematopoietic stem cell transplantation inform recommendations for follow-up. Expert Rev. Hematol. *4*, 437–452, quiz 453–454.

Biasco, L., Pellin, D., Scala, S., Dionisio, F., Basso-Ricci, L., Leonardelli, L., Scaramuzza, S., Baricordi, C., Ferrua, F., Cicalese, M.P.P., et al. (2016). In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. Cell Stem Cell *19*, 107–119.

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

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.

Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. *10*, 33.

Boettcher, S., Wilk, C.M., Singer, J., Beier, F., Burcklen, E., Beisel, C., Ventura Ferreira, M.S., Gourri, E., Gassner, C., Frey, B.M., et al. (2020). Clonal hematopoiesis in donors and long-term survivors of related allogeneic hematopoietic stem cell transplantation. Blood *135*, 1548–1559.

Boiteux, S., Coste, F., and Castaing, B. (2017). Repair of 8-oxo-7,8-dihydroguanine in prokaryotic and eukaryotic cells: Properties and biological roles of the Fpg and OGG1 DNA N-glycosylases. Free Radic. Biol. Med. *107*, 179–201.

Brem, R., Macpherson, P., Guven, M., and Karran, P. (2017). Oxidative stress induced by UVA photoactivation of the tryptophan UVB photoproduct 6-formylindolo[3,2-b]carbazole (FICZ) inhibits nucleotide excision repair in human cells. Sci. Rep. *7*, 4310.

Burns, S.S., and Kapur, R. (2020). Clonal Hematopoiesis of Indeterminate Potential as a Novel Risk Factor for Donor-Derived Leukemia. Stem Cell Reports *15*, 279–291.

Cameron, D.L., Baber, J., Shale, C., Papenfuss, A.T., Espejo Valle-Inclan, J., Besselink, N., Cuppen, E., and Priestley, P. (2019). GRIDSS, PURPLE, LYNX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. bioRxiv. https://doi.org/10.1101/781013.

Catlin, S.N., Busque, L., Gale, R.E., Guttorp, P., and Abkowitz, J.L. (2011). The replication rate of human hematopoietic stem cells in vivo. Blood *117*, 4460–4466.

Chen, H., Beardsley, G.P., and Coen, D.M. (2014). Mechanism of ganciclovir-induced chain termination revealed by resistant viral polymerase mutants with reduced exonuclease activity. Proc. Natl. Acad. Sci. USA *111*, 17462–17467.

Christopher, M.J., Petti, A.A., Rettig, M.P., Miller, C.A., Chendamarai, E., Duncavage, E.J., Klco, J.M., Helton, N.M., O'Laughlin, M., Fronick, C.C., et al. (2018). Immune Escape of Relapsed AML Cells after Allogeneic Transplantation. N. Engl. J. Med. *379*, 2330–2341.

Clark, C.A., Savani, M., Mohty, M., and Savani, B.N. (2016). What do we need to know about allogeneic hematopoietic stem cell transplant survivors? Bone Marrow Transplant. *51*, 1025–1031.

Collins, F.S., and Gottlieb, S. (2018). The next phase of human gene-therapy oversight. N. Engl. J. Med. *379*, 1393–1395.

Crumpacker, C.S. (1992). Mechanism of action of foscarnet against viral polymerases. Am. J. Med. *92* (2A), 3S–7S.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46* (D1), D794–D801.

De Ravin, S.S., Wu, X., Moir, S., Anaya-O'Brien, S., Kwatemaa, N., Littel, P., Theobald, N., Choi, U., Su, L., Marquesen, M., et al. (2016). Lentiviral hematopoietic stem cell gene therapy for X-linked severe combined immunodeficiency. Sci. Transl. Med. *8*, 335ra57.

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

Dunbar, C.E., High, K.A., Joung, J.K., Kohn, D.B., Ozawa, K., and Sadelain, M. (2018). Gene therapy comes of age. Science *359*, eaan4672.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47* (D1), D766–D773.

Gondek, L.P., Zheng, G., Ghiaur, G., DeZern, A.E., Matsui, W., Yegnasubramanian, S., Lin, M.T., Levis, M., Eshleman, J.R., Varadhan, R., et al. (2016). Donor cell leukemia arising from clonal hematopoiesis after bone marrow transplantation. Leukemia *30*, 1916–1920.

Griffiths, P., and Lumley, S. (2014). Cytomegalovirus. Curr. Opin. Infect. Dis. *27*, 554–559.

Hacein-Bey-Abina, S., Garrigue, A., Wang, G.P., Soulier, J., Lim, A., Morillon, E., Clappier, E., Caccavelli, L., Delabesse, E., Beldjord, K., et al. (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J. Clin. Invest. *118*, 3132–3142.

Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E., Hess, J.M., Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, L.Z., et al. (2016). Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell *164*, 538–549.

Hasaart, K.A.L., Manders, F., van der Hoorn, M.-L., Verheul, M., Poplonski, 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 haematopoiesis. Sci. Rep. *10*, 12991.

Howe, S.J., Mansour, M.R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., Brugman, M.H., Pike-Overzet, K., Chatters, S.J., de Ridder, D., et al. (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J. Clin. Invest. *118*, 3143–3150.

Husby, S., Favero, F., Nielsen, C., Sørensen, B.S., Bæch, J., Grell, K., Hansen, J.W., Rodriguez-Gonzalez, F.G., Haastrup, E.K., Fischer-Nielsen, A., et al. (2020). Clinical impact of clonal hematopoiesis in patients with lymphoma undergoing ASCT: a national population-based cohort study. Leukemia *34*, 3256–3268.

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.

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

Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, A.S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S.P., et al. (2019). A compendium of mutational signatures of environmental agents. Cell *177*, 821–836.e16.

Kuijk, E., Jager, M., van der Roest, B., Locati, M.D., van Hoeck, A., Korzelius, J., Janssen, R., Besselink, N., Boymans, S., van Boxtel, R., et al. (2020). The mutational impact of culturing human pluripotent and adult stem cells. Nat. Commun. *11*, 2493.

Lamm, N., Ben-David, U., Golan-Lev, T., Storchová, Z., Benvenisty, N., and Kerem, B. (2016). Genomic Instability in Human Pluripotent Stem Cells Arises from Replicative Stress and Chromosome Condensation Defects. Cell Stem Cell *18*, 253–261.

Lee-Six, H., Øbro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K., Belmonte, M., Osborne, R.J., Huntly, B.J.P., Martincorena, I., Anderson, E., et al. (2018). Population dynamics of normal human blood inferred from somatic mutations. Nature *561*, 473–478.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18–22.

Loeb, L.A., Essigmann, J.M., Kazazi, F., Zhang, J., Rose, K.D., and Mullins, J.I. (1999). Lethal mutagenesis of HIV with mutagenic nucleoside analogs. Proc. Natl. Acad. Sci. USA *96*, 1492–1497.

Lombard, D.B., Chua, K.F., Mostoslavsky, R., Franco, S., Gostissa, M., and Alt, F.W. (2005). DNA repair, genome stability, and aging. Cell *120*, 497–512.

Lu, R., Czechowicz, A., Seita, J., Jiang, D., and Weissman, I.L. (2019). Clonal-level lineage commitment pathways of hematopoietic stem cells in vivo. Proc. Natl. Acad. Sci. USA *116*, 1447–1456.

Lüdecke, D. (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. J. Open Source Softw. *3*, 772.

Maggs, D.J., and Clarke, H.E. (2004). In vitro efficacy of ganciclovir, cidofovir, penciclovir, foscarnet, idoxuridine, and acyclovir against feline herpesvirus type-1. Am. J. Vet. Res. *65*, 399–403.

Majhail, N.S., Tao, L., Bredeson, C., Davies, S., Dehn, J., Gajewski, J.L., Hahn, T., Jakubowski, A., Joffe, S., Lazarus, H.M., et al. (2013). Prevalence of hematopoietic cell transplant survivors in the United States. Biol. Blood Marrow Transplant. *19*, 1498–1501.

Mandai, M., Watanabe, A., Kurimoto, Y., Hirami, Y., Morinaga, C., Daimon, T., Fujihara, M., Akimaru, H., Sakai, N., Shibata, Y., et al. (2017). Autologous Induced Stem-Cell-Derived Retinal Cells for Macular Degeneration. N. Engl. J. Med. *376*, 1038–1046.

Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., Royo, R., Ziccheddu, B., Puente, X.S., Avet-Loiseau, H., et al. (2019). A practical guide for mutational signature analysis in hematological malignancies. Nat. Commun. *10*, 2969.

Morganella, S., Alexandrov, L.B., Glodzik, D., Zou, X., Davies, H., Staaf, J., Sieuwerts, A.M., Brinkman, A.B., Martin, S., Ramakrishna, M., et al. (2016). The topography of mutational processes in breast cancer genomes. Nat. Commun. *7*, 11383.

Mouhieddine, T.H., Sperling, A.S., Redd, R., Park, J., Leventhal, M., Gibson, C.J., Manier, S., Nassar, A.H., Capelletti, M., Huynh, D., et al. (2020). Clonal hematopoiesis is associated with adverse outcomes in multiple myeloma patients undergoing transplant. Nat. Commun. *11*, 2996.

Ortmann, C.A., Dorsheimer, L., Abou-El-Ardat, K., Hoffrichter, J., Assmus, B., Bonig, H., Scholz, A., Pfeifer, H., Martin, H., Schmid, T., et al. (2019). Functional Dominance of CHIP-Mutated Hematopoietic Stem Cells in Patients Undergoing Autologous Transplantation. Cell Rep. *27*, 2022–2028.e3.

Osorio, F.G., Rosendahl Huber, A., Oka, R., Verheul, M., Patel, S.H., Hasaart, K., de la Fonteijne, 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*, 2308–2316.e4.

Pasquini, M.C., Wang, Z., Horowitz, M.M., and Gale, R.P. (2010). 2010 report from the Center for International Blood and Marrow Transplant Research (CIBMTR): current uses and outcomes of hematopoietic cell transplants for blood and bone marrow disorders. Clin. Transpl. 87–105.

Passweg, J.R., Baldomero, H., Bader, P., Bonini, C., Cesaro, S., Dreger, P., Duarte, R.F., Dufour, C., Kuball, J., Farge-Bancel, D., et al. (2016). Hematopoietic stem cell transplantation in Europe 2014: more than 40 000 transplants annually. Bone Marrow Transplant. *51*, 786–792.

Piketty, C., Bardin, C., Gilquin, J., Gairard, A., Kazatchkine, M.D., and Chast, F. (2000). Monitoring plasma levels of ganciclovir in AIDS patients receiving oral ganciclovir as maintenance therapy for CMV retinitis. Clin. Microbiol. Infect. *6*, 117–120.

Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., van Hoeck, A., Wood, H.M., Nomburg, J., Gurjao, C., Manders, F., Dalmasso, G., Stege, P.B., et al.; Genomics England Research Consortium (2020). Mutational signature in colorectal cancer caused by genotoxic pks$^+$ E. coli. Nature *580*, 269–273.

Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210–216.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

Rosendahl Huber, A., Manders, F., Oka, R., and van Boxtel, R. (2019). Characterizing Mutational Load and Clonal Composition of Human Blood. J. Vis. Exp. *11* (149).

Scala, S., Basso-Ricci, L., Dionisio, F., Pellin, D., Giannelli, S., Salerio, F.A., Leonardelli, L., Cicalese, M.P., Ferrua, F., Aiuti, A., and Biasco, L. (2018). Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. Nat. Med. *24*, 1683–1690.

Seley-Radtke, K.L., and Yates, M.K. (2018). The evolution of nucleoside analogue antivirals: A review for chemists and non-chemists. Part 1: Early structural modifications to the nucleoside scaffold. Antiviral Res. *154*, 66–86.

Stratmann, S., Yones, S.A., Mayrhofer, M., Norgren, N., Skaftason, A., Sun, J., Smolinska, K., Komorowski, J., Herlin, M.K., Sundström, C., et al. (2021). Genomic characterization of relapsed acute myeloid leukemia reveals novel putative therapeutic targets. Blood Adv. *5*, 900–912.

Stunnenberg, H.G., and Hirst, M.; International Human Epigenome Consortium (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell *167*, 1145–1149.

Sun, J., Ramos, A., Chapman, B., Johnnidis, J.B., Le, L., Ho, Y.-J.J., Klein, A., Hofmann, O., and Camargo, F.D. (2014). Clonal dynamics of native haematopoiesis. Nature *514*, 322–327.

Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. *47* (D1), D941–D947.

Thompson, O., von Meyenn, F., Hewitt, Z., Alexander, J., Wood, A., Weightman, R., Gregory, S., Krueger, F., Andrews, S., Barbaric, I., et al. (2020). Low rates of mutation in clinical grade human pluripotent stem cells under different culture conditions. Nat. Commun. *11*, 1528.

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

Tomkova, M., Tomek, J., Kriaucionis, S., and Schuster-Böckler, B. (2018). Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. *19*, 129.

Weiner, J. (2017). riverplot: Sankey or Ribbon Plots. https://rdrr.io/cran/riverplot/.

Weissman, I.L. (2000). Stem cells: units of development, units of regeneration, and units in evolution. Cell *100*, 157–168.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer).

Xu, L., Wang, J., Liu, Y., Liangfu, X., Su, B., Mou, D., Wang, L., Liu, T., Wang, X., Zhang, B., et al. (2019). CRISPR-Edited Stem Cells in a Pateint with HIV and Acute Lymphocytic Leukemia. N. Engl. J. Med. *381*, 1240–1247.

Yamanaka, S. (2020). Pluripotent stem cell-based therapy - Promise and challenges. Cell Stem Cell *27*, 523–531.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD34-BV421, clone 561	BioLegend	Cat# 343609; RRID: AB_2561358
CD38-PE, clone HIT2	BioLegend	Cat# 303505; RRID: AB_314357
CD45RA-PerCp/Cy5.5, clone HI100	BioLegend	Cat# 304121; RRID: AB_893358
CD49f-PE/Cy7, clone GoH3	BioLegend	Cat# 313622; RRID: AB_2561705
CD90-APC, clone 5E10	BioLegend	Cat# 328113; RRID: AB_893440
Lineage(CD3/CD14/CD19/CD20/CD56)-FITC, clones UCHT1, HCD14, HIB19, HCD56)	BioLegend	Cat# 348701; RRID: AB_10644012
CD11c-FITC, clone 3.9	BioLegend	Cat# 301603; RRID: AB_314173
CD16-FITC, clone 3G8		Cat# 302005; RRID: AB_314205
Anti-phospho-γH2AX-FITC, clone, JBW301	Merk	Cat# 16-202A; RRID: AB_568825
Mouse IgG-FITC isotype control	Merk	Cat# 12-487; RRID: AB_436046
Biological samples		
HSCT donor bone marrow samples	Wilhelmina Children's Hospital and University Medical Center Utrecht	N/A
HSCT recipient blood samples, Sib1, Sib3	Wilhelmina Children's Hospital and University Medical Center Utrecht	N/A
HSCT recipient blood/bone marrow samples, Sib2, UCB1, UCB2, UCB3	Princess Máxima Center for Pediatric Oncology	N/A
Chemicals, peptides, and recombinant proteins		
Lymphoprep$^{\text{TM}}$ density gradient medium	Stem Cell Technologies	Cat# 07851
BD Vactutainer heparin tubes	BD Biosciences	Cat# 368480
Recombinant human thrombopoietin (TPO)	Preprotech	Cat# 300-18
Recombinant human stem cell factor (SCF)	Preprotech	Cat# 300-07
Recombinant human FLT3-L	Preprotech	Cat# 300-19
Recombinant human IL-6	Preprotech	Cat# 200-06
Recombinant human IL-3	Preprotech	Cat# 160-01
UM729	StemCell technologies	Cat# 72332
StemRegenin-1	StemCell technologies	Cat# 72342
Primocin	Invivogen	Cat# ant-pm-1
QIAamp DNA Micro kit	QIAgen	Cat# 56304
Formaldehyde solution	Sigma- Aldrich	Cat# F8775-25ML
Methanol	Sigma- Aldrich	Cat# 34860
Saponin	Millipore	Cat# 558255
BSA	Sigma- Aldrich	Cat# A7030-10G
HEPES	Thermo-Fisher Scientific	Cat# 15630106
PBS	Thermo-Fisher Scientific	Cat# 14190
DMSO	Sigma- Aldrich	Cat# D2438
EDTA	Sigma- Aldrich	Cat# T4049
NaN3	Sigma- Aldrich	Cat# 71289
FBS	Sigma- Aldrich	Cat# A4766801
Ganciclovir	Sigma- Aldrich	Cat# SML2346
Foscarnet sodium	Sigma- Aldrich	Cat# BP623

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Whole-genome sequence data from this article	This paper	European Genome-Phenome Archive (EGA; https://ega-archive.org/ega/home). Accession Number EGA:EGAS00001004926
Software and algorithms		
Whole genome sequencing read alignment and mutation calling pipeline		https://github.com/UMCUGenetics/IAP
SNV filtering pipeline		https://github.com/ToolsVanBox/SMuRF/
Indel filtering pipeline		https://github.com/ToolsVanBox/SMuRF/
R v3.6	R Core Team, 2020	https://www.R-project.org/
MutationalPatterns R package v3.0.1	Blokzijl et al., 2016	http://bioconductor.org/packages/3.12/bioc/html/MutationalPatterns.html
ggeffects R package v 0.14.2	Lüdecke, 2018	N/A
lme4 R package v1.1-21	Bates et al., 2015	N/A
randomForest R package v4.6-14	Liaw and Wiener, 2002	N/A
ggplot2 R package v3.2.1	Wickham, 2016	N/A
Riverplot R package v0.6	Weiner, 2017	N/A
dndscv R package v0.0.1.0		https://github.com/im3sanger/dndscv
Burrows-Wheeler Aligner v0.5.9 mapping tool	Li and Durbin, 2010	N/A
SAMTOOLS	Li et al., 2009	N/A
Structural variant caller grids-purple-linx	Cameron et al., 2019	https://github.com/hartwigmedical/gridss-purple-linx
Integrative Genomics Viewer (IGV)	Thorvaldsdóttir et al., 2013	N/A
Gencode v33	Frankish et al., 2019	https://www.gencodegenes.org/human/release_33.html
Encode	Davis et al., 2018	https://www.encodeproject.org/
Blueprint	Stunnenberg et al., 2016	http://dcc.blueprint-epigenome.eu/
Bedtools	Quinlan and Hall, 2010	https://bedtools.readthedocs.io/
UCSC liftOver		https://genome.ucsc.edu/cgi-bin/hgLiftOver

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ruben van Boxtel (r.van.boxtel@prinsesmaximacentrum.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The datasets generated during this study are available at EGA (https://www.ebi.ac.uk/ega/), accession number EGA:EGAS00001004926. Most of the scripts used during this study are available at https://github.com/ToolsVanBox/ and in the MutationalPatterns R package (see above). Other scripts are available upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

HSCT donor/recipient bone marrow and blood

Bone marrow cells of the HSCT donor were collected through the HSCT Biobank of the University Medical Center Utrecht. Peripheral blood and bone marrow of the HSCT recipients was obtained from the HSCT Biobank of the UMC Utrecht (SIB1 and SIB3), the Biobank of the Princess Máxima Center (CB1, CB2), or collected fresh by venipuncture into vacutainer tubes containing sodium heparin (SIB2, CB3, CB4, HAP1 donor and recipient, HAP2 donor and recipient). Details on samples and participants are depicted in Tables

S1 and S2. Informed consent was obtained from all participants and their caregivers. This study was approved by the Biobank Committee of the University Medical Center Utrecht (protocol number 18-231 and 19-737) and by the Medical Ethical Committee Utrecht (protocol number 19-243).

METHOD DETAILS

Cell isolation and flow cytometry

Mononuclear cells were isolated from whole blood and bone marrow using Lymphoprep density gradient separation (StemCell Technologies, Catalog# 07851). Single hematopoietic progenitor cells were sorted on a SH800S cell sorter (Sony), according to previously published methods (Osorio et al., 2018). The following combinations of cell surface markers were used to define cell populations : HSC: Lineage-CD34+CD38-CD45RA-CD90+CD11c-CD16- or Lineage-CD34+CD38-CD45RA-CD49f+CD11c-CD16-; MPP: Lineage-CD34+CD38-CD45RA-CD90-CD49f-CD11c-CD16-. Flow cytometry data were analyzed using the Sony SH800S Software (Sony). Polyclonal mesenchymal stromal cells (MSCs) were isolated from donor bone marrow samples by plating 0.5-1x10^6 donor cells in tissue-culture treated dishes in DMEM-F12 medium (GIBCO), supplemented with 10% fetal calf serum (FCS) and 1x Glutamax (GIBCO). Medium was replaced every 2-3 days to remove non-adherent cells. After 4-6 weeks, the adherent MSC fraction was isolated and used as a germline control.

FACS antibodies

The following antibodies were obtained from Biolegend and were used for HSPC isolation: CD34-BV421 (clone 561, 1:20; RRID AB_2561358); CD38-PE (clone HIT2, 1:50; RRID AB_314357), CD90-APC (clone 5E10, 1:200; RRID AB_893440), CD45RA-PerCP/Cy5.5 (clone HI100, 1:20; RRID AB_893358); CD49f-PE/Cy7 (clone GoH3, 1:100; RRID AB_2561705); CD16-FITC (clone 3G8, 1:100; RRID AB_314205); CD11c-FITC (clone 3.9, 1:20; RRID AB_314173), Lineage (CD3/CD14/CD19/CD20/CD56)-FITC (clones UCHT1, HCD14, HIB19, HCD56, 1:20; RRID AB_10644012). The following antibodies were obtained from Merk and were used for *γ-H2AX* expression staining: anti-phospho-histone H2A.X (Ser139) FITC conjugate (clone JBW301, 1:200; RRID AB_568825), mouse IgG FITC isotype control (1:200; RRID AB_436046).

Establishment of clonal HSPC cultures

HSPCs were index-sorted as single cells into round-bottom 384-well plates. Cells were cultured in StemSpan SFEM medium supplemented with SCF (100 ng/mL); FLT3-L (100 ng/mL); TPO (50 ng/mL); IL-6 (20 ng/mL) and IL-3 (10 ng/mL); UM729 (500 nM) and StemRegenin-1 (750 nM). After 3-6 weeks of culture at 37°C and 5% CO2, confluent colonies were collected for DNA isolation and sequencing.

Antiviral treatment of primary CD34+ cells *in vitro*

CD34+ cells were isolated from human umbilical cord blood by lymphoprep gradient separation and subsequent positive selection using the CD34+- UltraPure kit (Miltenyi Biotec) according to manufacturer's instructions. After an overnight incubation at 37°C, 5% O2 and 5% CO2, cells were treated with increasing concentrations of the following antiviral compounds: ganciclovir (Sigma Aldrich), foscarnet sodium (Sigma Aldrich), a combination of the two compounds or DMSO as vehicle control. Cells were incubated for 24 hours, after which DNA damage as assessed by γ-H2AX-staining and by WGS of clonally expanded cells.

For γ-H2AX-staining, 100,000-200,000 CD34+ cells were resuspended in permeabilization buffer containing 0.5% saponin, 0.5% BSA, 10mM HEPES, 140mM NaCl, 2.5mM CaCl2 in water, pH 7.4, sterile filtered. Anti-yH2A.X (Ser139) FITC (Merk) or Mouse IgG isotype antibody (X) were added to samples and cells were incubated for 20 min on ice. After staining, cells were washed with 0.1% saponin in PBS and resuspended in FACS buffer (1x PBS, 2%–5% FBS, 2mM EDTA, 2mM NaN_3) prior to flow cytometric analysis. For analysis of single-cell mutagenesis caused by antiviral treatment, CD34+ cells were sorted as single cells into flat-bottom 384-well plates (Greiner), using the same antibody mix and sorting strategy as for bone marrow and peripheral blood HSPCs. Cells were clonally expanded for 4-6 weeks, after which DNA was isolated (QIAamp DNA micro kit, QIAGEN) and sent for whole genome sequencing.

Analysis of γ-H2AX expression by flow cytometry

After drugs incubation, cells were harvested and washed with PBS. 100.000-200.000 CD34+ cells were resuspended in ice-cold fixative solution (2.5% formaldehyde and 0.93% methanol in sterile filtered PBS), incubated for 20 min at 4°C and transferred to a 96 well plate. Fixed samples were washed twice with PBS. Next, cells were resuspended in permeabilization buffer containing 0.5% saponin, 0.5% BSA, 10mM HEPES, 140mM NaCl, 2.5mM CaCl2 in water, pH 7.4, sterile filtered. Anti-yH2A.X (Ser139) FITC (Merk) or Mouse IgG isotype antibody (X) were added to samples and cells were incubated for 20 min on ice. After staining, cells were washed with 0.1% saponin in PBS and resuspended in FACS buffer (1x PBS, 2%–5% FBS, 2mM EDTA, 2mM NaN_3) prior to flow cytometric analysis on a Beckman Coulter CytoFLEX S.

Whole genome sequencing

DNA was isolated from the clonally expanded HSPCs using the DNeasy DNA Micro Kit (QIAGEN), according to the manufacturer's instructions. Libraries for Illumina sequencing were generated from 20-50 ng of genomic DNA using standard protocols (Illumina).

Samples were sequenced to 15-30x base coverage (2 × 150 bp) on an Illumina NovaSeq 6000 system. Sequence reads were mapped against the human reference genome (GRCh38) using the Burrows-Wheeler Aligner v0.7.5a mapping tool with settings 'bwa mem –c 100 –M' (Li et al., 2009). Sequence reads were marked for duplicates using Sambamba v0.6.8. Realignment was performed using the Genome Analysis Toolkit (GATK) version 3.8-1-0 (DePristo et al., 2011). A description of the complete data analysis pipeline is available at: https://gihub.com/UMCUGenetics/IAP.

Structural variants

Structural variant calling was done with the GRIDSS-purple-linx pipeline of the Hartwig Medical Foundation(Cameron et al., 2019). All resulting structural variants were checked by hand in the IGV(Thorvaldsdóttir et al., 2013) and false positive results were excluded. SVs could only be inspected of patients for which an MSC normal control was available.

Mutation calling and filtering

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 -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 post-processing filters as described (scripts available at: https://github.com/ToolsVanBox/SMuRF) (Blokzijl et al., 2016). Briefly, we considered variants at autosomal or X chromosomes without any evidence from a paired control sample if available (MSCs isolated from the same bone marrow); passed by VariantFiltration with a GATK phred-scaled quality score ≥ 100; a base coverage of at least 10X (30X samples) or 7X (15X samples) in the clonal 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; and absence of the variant in a panel of unmatched normal human genomes (BED-file available upon request). We additionally filtered base substitutions with a GATK genotype score (GQ) lower than 99 or 10 in clonal or paired control sample, respectively. For indels, we filtered variants with a GQ score lower than 99 in both clonal and paired control sample. In addition, for both SNVs and INDELs, we only considered variants with a variant allele frequency of 0.3 or higher for 30x coverage, and 0.15 or higher for 15x coverage in the clones to exclude *in vitro* accumulated mutations (Blokzijl et al., 2016; Jager et al., 2018). For patients for which no matched MSC, T cell or granulocyte control was available and clones were sequenced to 30x, we excluded mutations that were clonally present in all clones of the patient, or that were subclonally present in any clone of the patient. For patient CB3 no MSC control was available, and all clones were sequenced to 15x. For patient CB2 no control was available and three out of six cells were sequenced to 30x. For this sample, we applied the same filtering and in addition, we also filtered mutations that were not confidently absent in at least one sample. Lastly, we filtered out mutations that were clonal and/or failed QC in all, or all but one HSPC clones in that patient, as this suggests germline mutations that are missed in one or multiple cells due to low quality mapping or low coverage. Cells of these patients were re-sequenced to validate this approach.

Validation by re-sequencing

From leftover DNA of five HSPC clones included in this study, DNA libraries were constructed, sequenced to 15x and, processed as described above. 2 samples of patient CB2 that were previously sequenced to 30x and 3 samples of CB2 that were previously sequenced to 15x were included. Four out of these 5 harbored a high number of SBSA mutations. Mutations were deemed validated if the same mutations was found at a VAF of 0.15 or higher in the re-sequenced 15X sample.

HSPC mutation detection in bulk mature populations

For patient CB3, bulk B cells and bulk monocytes were sequenced to 30x and processed as described above, and the VAF of all mutations present in one or multiple HSPCs in this sample were assessed in these samples. All variants found in at least one reference allele were included in the analysis of Figure S4.

Baseline

For the baseline of age-related mutation accumulation in normal HSPCs, only autosomal chromosomes were considered. HSCT donor cells were used as part of the baseline. The number of SNVs or INDELs reported are normalized for the length of CALLABLE loci reported by GATK CallableLoci. For the slope estimation, the linear mixed-effects model was used to take donor dependency into account and the p values are indicated in the figures using lme4 package in R (Bates et al., 2015). The 0.95 confidence interval was calculated using the ggeffects package in R (Lüdecke, 2018). For comparison with the base line, we defined age of recipient HSPCs as the interval since birth, i.e., age of the donor added to the interval after HSCT. Plotting was done with ggplot2 in R (Wickham, 2016).

Assessment of C > A mutations in HSPC clones with increased mutation load

To statistically investigate the ratio of observed and expected mutations and the percentage of C > A mutations in the HSPC clones with an increased mutation load, a t test was applied from both data types to the HSCT donor and recipient clones that had an expected mutation load and the clones with an increased mutation load.

Mutational profile and signature analysis

We used an in-house developed R package (MutationalPatterns) (Blokzijl et al., 2018) to analyze mutational patterns. First, we extracted the 96-mutation profiles per sample. Then, we performed *de novo* mutational signature extraction on our data from HSCT donors and recipients, combined with healthy adult and pediatric tissue (Blokzijl et al., 2018; Osorio et al., 2018). The five extracted mutational patterns were compared to the COSMIC v3 signatures (Tate et al., 2019) together with our previously identified HSPC signature (Osorio et al., 2018) and based on their cosine similarities (> 0.9), three signatures were substituted by SBS signature 1, 5 and ‘HSPC’, resulting in SBS1, SBS5, HSPC, SBS18-like and SBSA. These signatures were subsequently refitted to the HSCT data, resulting in absolute contribution values. SBSA was compared to existing signatures (COSMIC v3; Tate et al., 2019) and signatures from Kucab et al. (2019) using cosine similarity of the 96-mutation profiles.

A modified version of the ‘‘calculate_lesion_segregation’’ function of MutationalPatterns was used to perform the Wald–Wolfowitz runs test for lesion segregation analysis, as described by Aitken et al. (2020), where the number of mutations and number of runs was pulled over samples in a group, before running the test. The baseline samples of individuals 40 years or older were used to ensure a sufficient number of mutations per sample. P values were corrected for multiple testing using Benjamini & Hochberg (FDR) correction (Benjamini and Hochberg, 1995).

Broader context of C > ApA mutations

To assess the broader context of C > ApA mutations of the SBSA signature, all C > ApA mutations were extracted from HSCT HSPCs with more than 70% contribution of SBSA and for the 875 and 260 μm potassium bromate signatures from Kucab et al. (2019). Next, for each sample the bases 10bp upstream (position −10) to 10 bp downstream (+10) of the mutated C (position 0) of these C > ApA mutations were extracted from the reference genome, and for each position the relative frequency of each of the 4 bases was calculated. The river plots were subsequently created for position −4 untill +4 by the R riverplot package v0.6 (Weiner, 2017).

Strand, genomic enrichment and replication bias analysis

We used the ‘‘‘‘mut_matrix_stranded’’ (with option ‘‘mode= ‘replication’ for replication direction), ‘‘strand_occurrences’’ and ‘‘strand_bias_test’’ functions of the in-house developed R package (MutationalPatterns) to determine transcription and replication strand bias (Blokzijl et al., 2018). We used the ‘‘genomic_distribution’’ and ‘‘enrichment_depletion_test’’ functions from the same package to analyze enrichment in genomic regions and early, mid and late replication regions. Gencode v33 was used to determine genomic regions (Frankish et al., 2019). Protein coding genes with the ‘‘appris_principal’’ tag were selected and the 100 bp around the 5′ end of genes was used as the transcription start site (TSS).

Processing of *in vitro* treated human umbilical cord blood cells

From cord blood sample CB22 (frozen), 1 ganciclovir treated clone, three foscarnet treated clones and three clones treated with both foscarnet and ganciclovir were sequenced. From cord blood sample CB25 (fresh) three untreated clones and three ganciclovir treated clones were sequenced. Library preparation, sequencing to 15X and data processing was performed as described above. In addition, only mutations observed in individual clones of a sample were considered to filter out *in vitro* acquired mutations.

Potential impact of mutational signatures

Calculating the probability of a mutation being caused by the signatures that contributed to that sample was done similar to Morganella et al. (2016). In short, the contributions of each signature to the sample were multiplied by the chance of each signature to induce a mutation of the mutation type and trinucleotide context of the driver mutation. These values were summed. The fraction that each signature contributed to the summed value was multiplied by 100 to get a probability in percentages.

The potential impact analysis from the new version of the MutationalPatterns package was used. In short, all the potential mutations in the coding sequence of 38 blood cancer driver genes were determined for each of the 96 mutation types. For each gene, the transcript with the longest combined coding sequence was used. For each mutation type the number of synonymous, missense and stop-gain mutations were then counted. A weighted sum over the 96 mutation types was then performed to determine the number of synonymous, missense and stop-gain mutations per signature, using the signature contributions as weights.

Random Forest

The “randomForest” function (option na.action = na.roughfix) of the randomForest R package v4.6-14 (Liaw and Wiener, 2002) was used to train the random forest. The input data for each single base substitution was as follows. (1) the −10:+10 nucleotide context, each position as a separate factor. (2) The distance to the nearest TSS and gene body (see above) and simple repeat calculated by “bedtools closest -d” (Quinlan and Hall, 2010). (3) The average Repliseq score from B lymphocytes obtained from ENCODE calculated by “bedtools intersect -wa -loj” (Wavelet-smoothed Signal bigWig, samples: Gm06990, Gm12801, Gm12812, Gm12813, Gm12878) (Davis et al., 2018). (4) The transcriptional strand bias calculated by comparing the DNA strand of the overlapping gene (“bedtools intersect -wa -loj”) with the strand of the mutated pyrimidine. (5) Gene expression of the overlapping gene (“bedtools intersect -wa -loj”). RNA-seq expression levels obtained from HSCs of the Blueprint DCC Portal (TPM value of “Transcription quantification (Genes)” files, samples: C002UUB1, C07002T1, C12001RP1) (Stunnenberg et al., 2016). (6) Reference and alternative allele. Results of bedtools intersect/closest was merged using “bedtools merge.” Mutations prediction was done by the “predict” function of the randomForest package. Mutation coordinates of reference genome hg38 were transferred to hg19 using UCSC’s liftOver.

Mutation datasets

The data of a knock-out of *OGG1* in the human neuroblastoma cell line CHP134 was courteously provided by Jan Molenaar (M.L. van den Boogaard, personal communication). Access to the WGS data of the 3668 Dutch metastases cohort from the Hartwig Medical Foundation can be requested at https://www.hartwigmedicalfoundation.nl/en/applying-for-data/. The CHIP and SCT databases were extracted from the supplemental information of the publications listed in Table 1 of Burns and Kapur (2020). The normal aging dataset used as control for the RF was extracted from the supplementary table of Lee-Six et al. (2018). The AML relapse data were obtained from Christopher et al. (2018) and Stratmann et al. (2021). Data on the post-HSCT neoplasms were obtained from Berger et al. (2018) and Gondek et al. (2016). The authors of Stratmann et al. (2021) provided us with all (unverified) genomic calls of the AML-relapses in their dataset that arose after HSCT. Upon suggestion of the authors, these were tested for COSMIC sequencing artifacts signatures. Each sample for which these artifacts contributed more than 20% were excluded from further analyses. Mutations were transferred to hg19 using UCSC’s liftOver. The aging CHIP dataset was obtained from Bick et al. (2020).

Construction of the phylogenetic lineage tree

To reconstruct the hematopoietic lineage tree of patient PMC11396 and HSCT recipients (Figure S4H), we compared the somatic base substitutions between whole-genome sequenced HSPC clones, and PMC11396’s primary ALL and tAML, using previously published data analysis pipelines (Osorio et al., 2018). To obtain base substitutions filtering was slightly altered compared to all other analyses to include mutations that were acquired during early embryonic development. When a control sample was available we included mutations with sub-clonal (VAF < 0.3) evidence in the paired control sample that were either clonally present or completely absent in all the clones. To still filter out germline mutations, only mutations that were confidently absent in at least one sample of a patient were included, only mutations for which all samples passed QC were considered, and mutations that were clonally present in all samples or subclonal in any samples were removed. All shared base substitutions were manually inspected. To summarize shared base substitutions, we created a binary mutation table. To construct the lineage trees, lineage distances were calculated using binary method, clones were hierarchically clustered using average method and plotted usingR.

QUANTIFICATION AND STATISTICAL ANALYSIS

Sample and mutation numbers are indicated in the figures. For estimation of the slope of age-related mutagenesis in normal HSPCs, a linear mixed-effects model was used, taking donor dependency into account. To assess statistical significance of lesion segregation the Wald- Wolfowitz runs test was performed. The statistical significance of transcription and replication strand bias was assessed by the Exact Poisson test (stats::poisson.test, R) and the statistical significance of genomic enrichment and depletion in regions of different replication timing was done by binomial testing (MutationalPatterns::binomial_test, R). The increase in percentage of C > A mutations in cells with an increased mutation burden was assessed with the Wilcoxon test. A Wilcoxon test was also used to compare γ-H2AX levels in *in vitro* treated cord blood cells. P values were Benjamini & Hochberg (FDR) corrected for multiple testing (R stats::p.adjust, option ‘method = “fdr”’).

ADDITIONAL RESOURCES

This study is registered in the Dutch Trial Register under study no. NL7585 (https://www.trialregister.nl).