Genomic Safety of Transplantation

Characterizing the Mutational Consequences of Treatment in Hematopoietic Stem Cells

Flavia Peci

Provided by thesis specialist Ridderprint, ridderprint.nl
Printing: Ridderprint
Layout and design: Jildou Hengst, persoonlijkproefschrift.nl
Cover Illustration : Benedetto Caselli

Genomic Safety of Transplantation:

Characterizing the Mutational Consequences of Treatment in Hematopoietic Stem Cells

De veiligheid van het genoom bij transplantatie:
Het karakteriseren van de mutationele gevolgen van behandeling in hematopoëtische stamcellen

(met een samenvatting in het Nederlands)

Proefschrift

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

donderdag 21 december 2023 des ochtends te 10.15 uur

door

Flavia Peci

geboren op 31 juli 1990
te San Benedetto del Tronto, Italië

Promotor:
Prof. dr. J.C. Clevers

Copromotoren:
Dr. R. van Boxtel
Dr. M.E. Belderbos

Beoordelingscommissie:
Dr. C. Janda
Prof. dr. M.M. Maurice
Prof. dr. M. van den Heuvel-Eibrink (voorzitter)
Prof. dr. J.H.E. Kuball
Prof. dr. H.G.P. Raaijmakers

Contents

General Introduction And Thesis Outline

1

Flavia Peci[1,2]

[1]*Princess Máxima Center for Pediatric Oncology, Utrecht, 3584 CS, The Netherlands*
[2]*Oncode Institute, Jaarbeursplein 6, 3521 AL Utrecht, The Netherlands*

Ever since the first allogeneic hematopoietic stem cell transplantation (HSCT) in 1957, advances in research and substantial changes in treatment regimens have greatly improved patient survival[1]. It is currently estimated that since 1957 to 2019 the number of transplanted patients has reached ~1.5 million worldwide[1]. Unfortunately, HSCT patients still have a lower life expectancy and suffers from mild to life-threatening side effects, such as actute and chronic graft versus host disease (GvHD), graft rejection, and infections[2]. Additional research is required to enhance transplantation regimens, minimize the toxicity of treatments, and attain disease-free survival. As the transplanted population's lifespan continues to extend, it becomes crucial to assess the influence of engrafted hematopoietic stem and progenitor cell (HSPC) clones and their long-term stability. It should be noted that HSPCs accumulate somatic mutations during life overtime, some of which may have neutral effects, while others could confer fitness advantages or disadvantages, leading to senescence or clonal expansion. There are several intriguing questions within the field of HSCT medicine and the long-term clinical impact on patients. For instance, it remains unclear what proportion of transplanted stem cells successfully engraft and thrive in the recipient over the long term, actively participating in hematopoietic regeneration. Additionally, it is important to investigate whether transplantation itself induces any genomic damage and how the transplanted system's clonal composition appears. Currently, the long-term engraftment potential and contribution to long-term multi-lineage hematopoiesis of human transplanted hematopoietic stem cells is still debated. Traditionally, most research into HSPC engraftment and regeneration post-HSCT has been performed using *in vivo* mouse models, using limiting dilution assays or artificial markers to track stem cells clones over time in syngeneic recipients[3,4]. However, advances in Next Generation Sequencing (NGS) and single cell technologies has improved HSPCs tracking methodologies, showing that somatic mutations can be used as clonal markers for tracing cells in human tissues[5,6].

Ageing affects the regenerative capacity of the hematopoietic system and increases the risk of blood cancer. On the cellular level, ageing results in telomere attrition, epigenetic changes, oxidative and/or replicative stress as well as the accumulation of somatic mutations[7]. Every cell of our body gradually acquires new somatic mutations with every cell division starting from when the zygote starts dividing[8]. By the age of 60, virtually all HSPCs of our body will have acquired a set of ~1000 somatic mutations[6], accounting for an average for a total of 30 million mutations in post-HSCT engrafting stem cells. The advent of next generation sequencing (NGS) technologies led to an improved understanding

of the impact of mutation accumulation in blood and other tissues in health and diseases[6]. As deep sequencing of bulk blood samples currently cannot fully recapitulate the clonal relationships between cells and does not allow for the detection of mutations present in a low proportion of blood cells, such as for example HSPCs, single cell sequencing analysis becomes pivotal. By applying single cell whole genome sequencing (WGS) to hematopoietic stem cells collected from the recipient and donor pairs, it is possible to perform data analysis to measure base substitutions at the single cell level as well as extract mutational signatures, detect chromosomal aberrations, and reconstruct developmental lineage trees[9]. This allows for the study of the human blood system in the recipient and directly compare this to the unperturbed condition in the donor. However, despite the successful applications of the ***in vitro*** clonal expansion method, limitations remains. Infact, single cell WGS relies on the clonal expansion capacity of stem cells as the amount of DNA from one cell is not enough for generating sequencing libraries. Whilst ***in vitro*** clonal expansion is possible for healthy HSPCs, clonal expansion of genetically impaired Fanconi Anemia (FA) HSPCs, and therefore their genome amplification with this method, is challenging, as the cell genome is compromised by genetic aberrations and chromosomal instability. Therefore, for genome analysis of cells such as FA HSPCs, who lack ***in vitro*** clonal expansion capacity, other methods that allow the genome amplifications are needed. One possibility is offered by a new technique called primary template amplification (PTA)[10]. Additionally, pre-HSCT conditioning through high-dose chemotherapy and/or irradiation can cause substantial short- and long-term toxicity to the BM niche[11]. Damage to the niche may also impair HSC function as well as hematopoietic regeneration post-HSCT and predispose to HSCT-related morbidity and mortality[11]. HSCT patients are also at increased risk of developing bacterial, fungi and viral infection after transplantation, as the recipient immune system is suppressed during treatment[11]. To treat opportunistic infections, antiviral nucleoside analog (NA) drugs are often use in the management of viral re-activation after HSCT[12]. Although the toxicity profile of these drug has been tested ***in vitro*** via functional assays, the mutagenic and carcinogenic effects of these drugs on the transplanted recipient stem cells is unknown.

HEMATOPOIETIC STEM CELL TRANSPLANTATION

Hematopoietic stem cell transplantation (HSCT) is a medical procedure that involves the infusion of hematopoietic stem and progenitor cells (HSPCs) from a matched donor into a recipient with hematological malignancies, immunodeficiencies, or genetic disorders, such as Fanconi Anemia (FA)[13]. The

success rate and overall survival of transplanted patients have increased in the past decade, thanks to reduced intensity of the pre-conditioning regimens, better supportive care, expanded graft source options, such as the possibility of obtaining transplantable hematopoietic cells from multiple sources, and continuous monitoring of HSCT survivors[14]. However, preventing and treating failed engraftment, managing graft versus host disease (GvHD), and maintaining an immunocompetent state of the recipient remain major challenges in transplantation medicine[15]. Furthermore, due to the transiently deprived immunological state until donor cell recovery, HSCT recipients are also at risk of bacterial, viral, and fungal infections. Furthermore, HSCT survivors often experience late adverse effects of the treatment, which can have detrimental effects on the entire body later in life. These late effects can include cardiovascular, pulmonary, and endocrine system dysfunctions, bone diseases, solid cancer, neurophysiological effects, and premature death[15,16]. While some of these effects may be due to the toxicity of chemotherapy and pre-conditioning regimens, other late effects such as clonal hematopoiesis in the recipient may be directly related to the transplanted blood product. Indeed, clonal hematopoiesis is associated with an increased risk on hematological cancers and coronary heart disease[17,18]. The clonal dynamics by which the donor's HSPCs regenerate the recipient's blood and the mutational consequences of transplantation on the aging of stem cells are still unclear. Somatic mutations are accumulated linearly over time in blood stem cells at the rate of ~15-17 mutations per year[6,19]. However, the mutational consequences of HSCT in recipients have never been measured. Also, the number of stem cells engrafting and contributing to the recipient's blood production is an ongoing debate[20]. Recipients may be at an increased risk of clonal hematopoiesis due to limited number of transplanted stem cells[21] or the increase of somatic mutations, which might give rise to mutated clones that are better adapted to the recipient's bone marrow. Clonal hematopoiesis is a phenomenon that occurs in a third of the general population when they age and is defined by the expansion of a hematopoietic stem cell due to specific somatic driver mutations[22]. Interestingly, the presence of clonal hematopoiesis can be linked to cardiovascular disease and can precede the development of myeloid malignancies years before diagnosis[23]. The hypothesis that donor clonal hematopoiesis could play a role in survivors health status has lead to multiple investigations[24,25] (Trial no. NCT04689750 and no. NL9585). Therefore, ensuring the genomic safety of the transplanted product, particularly in pediatric patients, is crucial for improving quality of life and survival, especially in pediatric cancer patients. **Chapter 3**

of this thesis descibes a detailed investigation of the genomic consequences of HSCT.

THE GENETIC FRAGILITY OF FANCONI ANEMIA HEMATOPOIETIC STEM CELLS

As above mentioned, HSCT is also used in the treatment of a variety of genetic disorders, such as Fanconi Anemia (FA). FA is a rare autosomal recessive disease characterized by congenital anomalies, increased cancer risk, pancytopenia and progressive bone marrow failure (BMF)[26]. From the molecular point of view, FA is a chromosome instability syndrome, characterized by mutations to one of the 23 genes involved in the FA pathway, a DNA repair pathway that reconizes damage caused by interstrand crosslinks[27]. Various research efforts have tried to pinpoint the mechanism driving the onset of BMF in FA[28], however, the nature of this condition seems to be multifaceted and difficult to tackle due to the fragile nature of FA HSPCs, limitations related to single cell sequencing applications, sample frequency and handling. Previous studies have shown that FA HSPCs have hyperactivated p53/p21 as well as MYC pathways with implications on replication stress and DNA damage accumulation which leads to progressive BMF, clonal hematopoiesis and hematological malignancies[28–30]. Due to their intrinsic genomic fragility, experimental modelling of primary FA-HSPCs in culture systems has been challenging. Recent efforts have focused on modelling overt leukemia phenotype and the generation of pre-leukemia FA mice[31]. Although rodents have been successfully used in the study of different hematological malignancies, they pose a limitation in the study of FA. In contrast to humans, mice exibit some but not all of the developmental and hematological charateristics of human FA patients[32]. Whilst mouse studies offer overall valuable insights, there is a preference for utilizing *in vitro* experimental methods with patient materials to enhance our understanding even further. As described above, single-cell WGS provides unique opportunities to understand the role of somatic mutation accumulation in FA disease progression. However, attempts to investigate FA-HSPC genomes by using single cell WGS have been hampered by the incapacity of cells to clonally expand *in vitro*. Although methods, such as single molecule duplex sequencing can be applied to non-dividing cells with very high accuracy[33], for some other techniques such as Primary Template-Directed Amplification (PTA), low accuracy in mutation detection remains a challenge. In **chapter 6**, we use PTA to characterize the genome of single FA HSPCs. By analysing the WGS data with a new bioinformatic pipeline PTA Analysis Tool (PTATO), we improved accuracy of PTA and observed that HSPCs from FA patients have a normal mutational burden and high number of deletions[10].

GENOME ANALYSIS OF HEMATOPOIETIC STEM CELLS BY SINGLE CELL SEQUENCING

Over the last few decades, the study of somatic mutations has been revolutionized by the improvement of next generation sequencing (NGS) techniques, which enabled WGS analysis. Analysing genomes with WGS enables to reach genome-wide coverage, in contrast to other methods such as targeted sequencing. As each HSPCs harbors a unique set of somatic mutations, single cell genome analysis is required in contrast to bulk DNA sequencing to pick up cell-specific somatic mutations. Besides identifying mutations that can driver carcinogenesis (i.e., driver mutations), mutations can also be used to identify the processes that caused mutagenesis. For this, signature analysis has been used as analysis tools that contributed to investigate mutational etiology in different tissues[34,35]. As signatures exists for all type of mutations present across genomes, it is possible to classify them because of their characteristic features, such as type and occurrence. This results in very specific patterns that reflect the activity of individual mutagenic processes[36]. Using mutational signatures (MS) analysis, the processes underlying lifelong mutation accumulation in HSPCs were reflected by SBS1, SBS5 and HSPC[6,37]. SBS1 and SBS5, and HSPC namely clock-like signatures, reflect stem cells aging process in the hematopoietic compartment. The etiology for SBS1 is known, namely spontaneous deamination of 5-methylcytosyne; however, the causes of SBS5 and HSPC are currently under investigation[34,37]. By assessing somatic mutations that are shared between different cells of the same donor, we can study the pattern of sharing mutations among cells. In this respect, the phylogeny of blood can be reconstructed using somatic mutations as unique barcodes to find shared ancestral cells by looking at shared mutations. The study of non-malignant tissue by WGS lineage tracing has revealed that polyclonality is a feature of healthy tissues. Of note, high clonal diversity is maintained in human hematopoiesis until 65 years of age[38]. In contrast, loss of polyclonality is observed in blood over the age of 75. According to recent investigations, it is estimated that a population of approximately 50,000-250,000 HSPCs contributes to blood production in healthy individuals[5]. However, when it comes to HSCT transplanted patients, little is known on HSPCs ageing and blood clonality. Although it is an important factor to consider, the mutational processes, such as consequences of replication stress, that newly transplanted HSPCs will undergo to regenerate the recipient hematopoietic system remain partly unknown, and currently under study[20].

DRUG TREATMENT EFFECTS IN HEMATOPOIETIC STEM CELL TRANSPLANTATION

Prior to HSCT, patients undergo conditioning treatment which involves the use of high-dose chemotherapy, monoclonal antibodies, and/or irradiation to

eradicate the primary cause of disease and facilitate donor cell engraftment by avoiding graft rejection via immunosuppression[39,40]. Although, this pre-transplant treatment is a necessary step to eliminate cancer cells, healthy cells within the bone marrow niche will also be exposed to it. The short and long-term toxicity associated with the conditioning regimen on the bone marrow niche is not entirely understood on a molecular level, and since the niche is an indispensable microenvironment, which support hematopoietic regeneration post-HSCT and dictates stem cell fate, it is critical to investigate the impact of conditioning on bone marrow niche cells[11,41]. It is noteworthy that changes occurring in the bone marrow niche of the recipient, encompassing non-hematopoietic cells that offer support to transplanted HSPCs, have the potential to influence engraftment and hinder the success of HSCT. **Chapter 2** reviews the effects of conditioning regimens on the bone marrow niche and proposes potential strategies to prevent or repair the resulting damage. The ultimate goal is to improve hematopoietic recovery and enhance the outcome of hematopoietic stem cell transplantation. Additionally, the success of HSCT depends on factors such as human leukocyte antigen (HLA) marker and serostatus compatibility between donor and recipient. A close HLA match is associated with better outcomes in terms of engraftment and rejection prevention as well as serostatus compatibility, the most important determinant of Cytomegalovirus (CMV) reactivation after transplant. Of note, infections pre- and post-engraftment of the transplanted HSPCs play a critical role. Various viruses can afflict HSCT patients, including CMV, varicella-zoster virus (VZV), herpes simplex virus (HSV), and adenoviruses. To treat such infections, HSCT patients typically receive pro-drugs like valacyclovir, famciclovir, and valganciclovir for anti-herpetic treatment[42]. However, the genomic safety profile of these drugs is currently uncertain[43,44]. The majority of antiviral drugs currently approved for the treatment of viral infections im immunosuppressed patients are nucleoside analogs (NA). Because of their mechanism of action, antiviral NAs can be mutagenic to non-infected cells. An example of this is ganciclovir (GCV), a guanosine analog which has shown a high toxicity profile and is used for CMV reactivation post-HSCT. The mutagenic and carcinogenic potential of GCV to healthy cells has been documented[9,45] . Recently a debate was sparked by the scientific community on the human genetic risk on the use of another antiviral NA named Molnupiravir, which is a recent FDA approved treatment for SARS-CoV-2 virus[46]. Moreover, recent studies have highlighted the mutagenicity and carcinogenicity of GCV in transplanted patients treated for CMV by looking at mutational profiles from of 121,771 patient samples in the GENIE and FM cohorts[45]. In **Chapter 3** we present a study in which we show that

GCV is mutagenic to transplanted HSPCs in cancer survivors and have high chances of causing driver mutations. Furthermore, the lack of consistency in current genotoxic screening methods for NA antiviral compounds used in clinics warrants further investigation.

OUTLINE OF THE THESIS

With the work described in this thesis, we aimed to characterize the genomic safety of HSCT, a therapeutic approach currently used in more than 40.000 patients worldwide each year as treatment for hematological malignancies, immunodeficiencies and genetic disorders.

In this general introduction (**Chapter 1**), we provided a comprehensive overview on the different aspects of HSCT, with a focus on the genomic safety.

Chapter 2 reviews how the toxicity of conditioning treatment in HSCT affects the recipient's bone marrow niche. We conducted a detailed analysis of the effect of conditioning on each cell type, revealing that it has an overall detrimental impact on the bone marrow niche. To promote hematopoietic recovery, further research is needed to understand how this damage can be prevented.

In **Chapter 3**, we conducted a study to accurately measure the mutational consequences of allogeneic HSCT in a cohort of 9 pediatric cancer patients who underwent successful engraftment after transplantation for hematological malignancy[9]. To achieve this, we used single-cell WGS to analyze the genome of HSPCs collected from donors and recipients at the same time, allowing us to investigate the hematopoietic system in both settings. We examined somatic mutations to determine the genomic age of the transplanted stem cells and performed mutational signature analysis to identify which processes were active in the cell's genome before and after transplantation. Our findings showed that, overall, allogeneic HSCT did not induce increased mutagenesis in the transplanted stem cells up to 2 years post-HSCT. However, we did observe an increased mutation load and unique mutational signature in some patients, which we validated *in vitro* and attributed to the antiviral NA drug ganciclovir. The method we used for *in vitro* validation is described in detail in **Chapter 4**[47]**.**

The described method allows for the investigation of the genotoxic effects of antiviral NA compounds in human HSPCS. For this protocol, CD34+ cells

are enriched from fresh donor umbilical cord blood and are then exposed to varying concentrations of a chosen compound to determine the IC40-IC60 concentration. Human umbilical cord blood-derived HSPCs offer an ideal cell source for mutational analysis due to their low mutation background, as these cells are still very young[47]. In **Chapter 5**, we developed a study as follow up of the work described in **Chapter 3**. We used single-cell WGS to examine the genomes of clonally expanded stem cells that were previously treated with antiviral compounds. The results showed that most of the compounds did not increase mutagenesis in HSPCs and are therefore considered safe. However, 5 of the 15 compounds in the list were found to significantly increase mutagenesis and alter the genome of HSPCs.

While single-cell WGS methods have become increasingly popular due to their success and versatility, limitations still hamper this technique. New ways for WGA are needed when clonal expansion of cells with standard culture methods is not possible, such as with differentiated cells or cells affected by an inherited genetic disorder like FA-HSPCs. In **Chapter 6**, we focus on the genomic consequences of FA in HSPCs and increase our understanding of the causes of bone marrow failure in FA. Here, we investigate freshly isolated FA-HSPCs with genome-wide coverage using a unique technique called PTA and a new bioinformatic workflow named PTATO[10]. The analysis revealed that FA-HSPCs are characterized by a normal mutation burden and an increased number of deletions. This finding is just one example of how new WGA techniques can advance our understanding of disease biology.

In **Chapter 7**, we provide a summary and broader discussion of our findings. Additionally, we present recommendations for future research.

ACKNOWLEDGEMENTS

I would like to thank Ruben van Boxtel and Mirjam Belderbos for providing feedback.

REFERENCES

1. Niederwieser, D. ***et al.*** One and a half million hematopoietic stem cell transplants: continuous and differential improvement in worldwide access with the use of non-identical family donors. ***Haematologica* 107**, 1045–1053 (2021).
2. Bhatia, S. Cause-specific late mortality after allogeneic stem cell transplantation. ***Hematology* 2019**, 626–629 (2019).
3. Vanuytsel, K. ***et al.*** Multi-modal profiling of human fetal liver hematopoietic stem cells reveals the molecular signature of engraftment. ***Nat Commun* 13**, 1103 (2022).
4. Omer-Javed, A. ***et al.*** Mobilization-based chemotherapy-free engraftment of gene-edited human hematopoietic stem cells. ***Cell* 185**, 2248-2264.e21 (2022).
5. Lee-Six, H. ***et al.*** Population dynamics of normal human blood inferred from somatic mutations. ***Nature* 561**, 473–478 (2018).
6. Osorio, F. G. ***et al.*** Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. ***Cell Rep* 25**, 2308-2316.e4 (2018).
7. Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in the ageing process. ***Nature* 592**, 695–703 (2021).
8. Park, S. ***et al.*** Clonal dynamics in early human embryogenesis inferred from somatic mutation. ***Nature* 597**, 393–397 (2021).
9. de Kanter, J. K. ***et al.*** Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. ***Cell Stem Cell* 28**, (2021).
10. Middelkamp, S. ***et al.*** Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox. ***bioRxiv*** (2023).
11. Peci, F. ***et al.*** The cellular composition and function of the bone marrow niche after allogeneic hematopoietic cell transplantation. ***Bone Marrow Transplant* 57**, 1357–1364 (2022).
12. Annaloro, C. ***et al.*** Viral Infections in HSCT: Detection, Monitoring, Clinical Management, and Immunologic Implications. ***Front Immunol* 11**, (2021).
13. Snowden, J. A. ***et al.*** Indications for haematopoietic cell transplantation for haematological diseases, solid tumours and immune disorders: current practice in Europe, 2022. ***Bone Marrow Transplant* 57**, 1217–1239 (2022).
14. Wauben, B. ***et al.*** Assessing long-term effects after stem cell transplantation: design of the MOSA study. ***J Clin Epidemiol* 148**, 10–16 (2022).
15. Inamoto, Y. & Lee, S. J. Late effects of blood and marrow transplantation. ***Haematologica* 102**, 614–625 (2017).
16. Diesch-Furlanetto, T. ***et al.*** Late Effects After Haematopoietic Stem Cell Transplantation in ALL, Long-Term Follow-Up and Transition: A Step Into Adult Life. ***Front Pediatr* 9**, (2021).
17. Xie, M. ***et al.*** Age-related mutations associated with clonal hematopoietic expansion and malignancies. ***Nat Med* 20**, 1472–1478 (2014).
18. Jaiswal, S. ***et al.*** Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. ***New England Journal of Medicine* 377**, 111–121 (2017).
19. Mitchell, E. ***et al.*** Clonal dynamics of haematopoiesis across the human lifespan. ***Nature* 606**, 343–350 (2022).
20. Campbell, P. ***et al.*** Clonal dynamics after allogeneic haematopoietic cell transplantation using genome-wide somatic mutations. doi:https://doi.org/10.21203/rs.3.rs-2868644/v1.
21. Warren, J. T. & Link, D. C. Clonal hematopoiesis and risk for hematologic malignancy. ***Blood*** (2020) doi:10.1182/blood.2019000991.

22. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nat Genet* **54**, 1155–1166 (2022).
23. Desai, P. *et al.* Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat Med* **24**, 1015–1023 (2018).
24. Burns, S. S. & Kapur, R. Clonal Hematopoiesis of Indeterminate Potential as a Novel Risk Factor for Donor-Derived Leukemia. *Stem Cell Reports* **15**, 279–291 (2020).
25. Boettcher, S. *et al.* Clonal hematopoiesis in donors and long-term survivors of related allogeneic hematopoietic stem cell transplantation. *Blood* **135**, 1548–1559 (2020).
26. Rageul, J. & Kim, H. Fanconi anemia and the underlying causes of genomic instability. *Environ Mol Mutagen* **61**, 693–708 (2020).
27. Smogorzewska, A. Fanconi Anemia: A Paradigm for Understanding DNA Repair During Replication. *Blood* **134**, SCI-32-SCI-32 (2019).
28. Ceccaldi, R. *et al.* Bone Marrow Failure in Fanconi Anemia Is Triggered by an Exacerbated p53/p21 DNA Damage Response that Impairs Hematopoietic Stem and Progenitor Cells. *Cell Stem Cell* **11**, 36–49 (2012).
29. Rodríguez, A. *et al.* MYC Promotes Bone Marrow Stem Cell Dysfunction in Fanconi Anemia. *Cell Stem Cell* **28**, 33-47.e8 (2021).
30. Sebert, M. *et al.* Clonal hematopoiesis driven by chromosome 1q/MDM4 trisomy defines a canonical route toward leukemia in Fanconi anemia. *Cell Stem Cell* **30**, 153-170.e9 (2023).
31. Cerabona, D., Sun, Z. & Nalepa, G. Leukemia and chromosomal instability in aged Fancc-/- mice. *Exp Hematol* **44**, 352–357 (2016).
32. Parmar, K., D'Andrea, A. & Niedernhofer, L. J. Mouse models of Fanconi anemia. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **668**, 133–140 (2009).
33. Bae, J. H. *et al.* Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. *Nat Genet* **55**, 871–879 (2023).
34. Petljak, M. *et al.* Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature* **607**, 799–807 (2022).
35. Thatikonda, V. *et al.* Comprehensive analysis of mutational signatures reveals distinct patterns and molecular processes across 27 pediatric cancers. *Nat Cancer* **4**, 276–289 (2023).
36. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
37. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402–1407 (2015).
38. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
39. Bredeson, C. N. *et al.* Outcomes following HSCT Using Fludarabine, Busulfan, and Thymoglobulin: A Matched Comparison to Allogeneic Transplants Conditioned with Busulfan and Cyclophosphamide. *Biology of Blood and Marrow Transplantation* **14**, 993–1003 (2008).
40. Andersson, B. S. *et al.* Clofarabine ± Fludarabine with Once Daily i.v. Busulfan as Pretransplant Conditioning Therapy for Advanced Myeloid Leukemia and MDS. *Biology of Blood and Marrow Transplantation* **17**, 893–900 (2011).
41. Fröbel, J. *et al.* The Hematopoietic Bone Marrow Niche Ecosystem. *Front Cell Dev Biol* **9**, (2021).

42. Jancel, T. & Penzak, S. R. Antiviral Therapy in Patients With Hematologic Malignancies, Transplantation, and Aplastic Anemia. *Semin Hematol* **46**, 230–247 (2009).
43. Wutzler, P. & Thust, R. Genetic risks of antiviral nucleoside analogues – a survey. *Antiviral Res* **49**, 55–74 (2001).
44. Jiang, L. ***et al.*** Genetic Evidence for Genotoxic Effect of Entecavir, an Anti-Hepatitis B Virus Nucleotide Analog. *PLoS One* **11**, e0147440 (2016).
45. Fang, H. ***et al.*** Ganciclovir-induced mutations are present in a diverse spectrum of post-transplant malignancies. *Genome Med* **14**, 124 (2022).
46. Waters, M. D., Warren, S., Hughes, C., Lewis, P. & Zhang, F. Human genetic risk of treatment with antiviral nucleoside analog drugs that induce lethal mutagenesis: the special case of molnupiravir. *Environ Mol Mutagen* **63**, 37–63 (2022).
47. Rosendahl Huber, A. ***et al.*** Whole-genome sequencing and mutational analysis of human cord-blood derived stem and progenitor cells. *STAR Protoc* **3**, 101361 (2022).

2

The cellular composition and function of the bone marrow niche after allogeneic hematopoietic cell transplantation

Flavia Peci*[1,3], Linde Dekker*[1], Anna Pagliaro[1], Ruben van Boxtel[1,3], Stefan Nierkens[1,2], Mirjam Belderbos[1]

[1]*Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands*
[2]*Center for Translational Immunology, University Medical Center Utrecht, Utrecht, The Netherlands*
[3] *Oncode Institute, Utrecht, The Netherlands*

* These authors contributed equally to this work

Bone Marrow Transplant, DOI: 10.1038/s41409-022-01728-0

Abstract

Allogeneic hematopoietic cell transplantation (HCT) is a potentially curative therapy for patients with a variety of malignant and non-malignant diseases. Despite its life-saving potential, HCT is associated with significant morbidity and mortality. Reciprocal interactions between hematopoietic stem cells (HSCs) and their surrounding bone marrow (BM) niche regulate HSC function during homeostatic hematopoiesis as well as regeneration. However, current pre-HCT conditioning regimens, which consist of high-dose chemotherapy and/or irradiation, cause substantial short- and long-term toxicity to the BM niche. This damage may negatively affect HSC function, impair hematopoietic regeneration after HCT and predispose to HCT-related morbidity and mortality. In this review, we summarize current knowledge on the cellular composition of the human BM niche after HCT. We describe how pre-HCT conditioning affects the cell types in the niche, including endothelial cells, mesenchymal stromal cells, osteoblasts, adipocytes and neurons. Finally, we discuss therapeutic strategies to prevent or repair conditioning-induced niche damage, which may promote hematopoietic recovery and improve HCT outcome.

Key words: Hematopoietic stem cell transplantation, bone marrow transplantation, stem cell niche, graft failure

Background

Allogeneic hematopoietic cell transplantation (HCT) is a potentially curative treatment for patients suffering from hematologic malignancies, red blood cell disorders, bone marrow failure, severe immune deficiency and certain metabolic disorders. About 20,000 and 40,000 allogeneic transplants are performed annually in Europe and the United States, respectively, and numbers are increasing[1,2]. However, 5-10% of HCT recipients experience graft failure, which is often fatal[3]. Furthermore, poor graft function affects up to 20% of HCT recipients and predisposes to infections, viral reactivations, bleeding complications, relapsed malignancy, and overall mortality[4].

Successful HCT requires depletion of the recipient's blood and immune system, followed by administration of donor hematopoietic stem cells (HSCs) which home and engraft the recipient's BM and reconstitute all the blood cell lineages. Depletion of the recipient's blood and immune system is achieved by pre-HCT conditioning, which consists of combinations of chemotherapy, radiotherapy and lymphodepleting agents. Post-transplant hematopoietic recovery typically occurs in phases: while innate immune cells and thrombocytes generally recover within weeks after HCT, complete reconstitution of adaptive immunity can take months to even years[5]. The slow reconstitution of the adaptive immune system is a result of ineffective thymic recovery, due to damage to the thymus by pre-HCT conditioning. Of note, T cells reconstitution consists in two phases: first, homeostatic proliferation of T cells from the graft; second, recovery of the thymus and thymic output. Although anti-Thymocyte Globulin (ATG) treatment can affect both stages, the homeostatic T cell proliferation is mainly impacted. Moreover, aGvHD and cGvHD can decrease thymic output, as reviewed by Velardi et al (2021)[6]. Overall, the dynamics of post-transplant hematopoietic and immune reconstitution is one of the most important determinants of HCT-related complications and survival[7].

Host HSCs, as well as transplanted HSCs, require support of a specialized bone marrow (BM) microenvironment, known as the "niche". The concept of a niche was first introduced in 1978 by Schofield, who postulated that the fate of a stem cell is dictated by the environment in which it resides[8]. Reciprocal interactions between HSCs and their niche regulate HSC quiescence, self-renewal, proliferation, differentiation, mobilization and homing[9]. Conditioning-induced niche damage also involves non-hematopoietic cells[10,11], which may

further affect hematopoietic recovery after HCT[9,12] and predispose to prolonged cytopenia, HSC non-engraftment and poor graft function[11,13].

Recent developments in single-cell sequencing and imaging have greatly improved our understanding of the cellular composition of the BM niche[14,15]. These studies uncovered a great level of complexity in the cellular and molecular constituents of the BM niche, as well as in the mechanisms by which they regulate HSC behaviour. However, thus far, most of these studies have been performed in mice, and studies in humans are only beginning to appear.

Here, we review current knowledge on the BM niche in the context of HCT. We summarize the effects of pre-HCT conditioning on each of the distinct BM niche cell types and on the mechanisms by which they support post-HCT hematopoietic regeneration. Furthermore, we discuss strategies to prevent or treat conditioning-induced niche damage, which may ultimately contribute to improved HCT outcome.

Architecture of the bone marrow niche

BM is a cell-dense, semi-solid tissue localized in the central (or medullary) cavities of axial and long bones. The BM is highly vascularized by an abundant heterogeneous network of blood vessels, which serves to supply nutrients, oxygen and signaling molecules, while removing waste products. Nevertheless, BM is a relatively hypoxic microenvironment [16], critically regulating HSC metabolism and quiescence; high reactive oxygen species (ROS) levels promote HSC differentiation and mobilization, whereas low levels of ROS promote HSC quiescence, self-renewal and long-term repopulating potential[17,18]. The BM niche vasculature is supported by an extensive network of multipotent mesenchymal stromal cells (MSCs), which can give rise to osteoblasts, chondrocytes and adipocytes[19,20]. In addition, bone as well as BM are highly innervated by autonomic and nociceptive nerve fibers and associated Schwann cells[21]. Below, we will discuss how these BM niche populations may be influenced by transplantation procedure and the subsequent effect on hematopoietic recovery after HCT.

Niche cells and their impact on hematopoietic recovery after HCT

Endothelial cells

Endothelial cells (ECs) form a monolayer that constitutes the inner lining of blood vessels and facilitate blood flow, enable exchange of nutrients and waste products, and regulate vascular tone and blood coagulation. Based on their localization within the BM vasculature, ECs can be classified as arteriolar endothelial cells (AECs) or sinusoid endothelial cells (SECs)[22] that differ in signaling molecules and modulation of the microenvironment, thus establishing distinct vascular niches that can instruct HSCs [22,23]. AECs are part of arteriolar vessels with low plasma penetration and maintain a relatively hypoxic environment[22,24]. They are a major source of netrin-1, which, through interaction with its receptor neogenin-1, serves to maintain HSC quiescence and self-renewal[25]. Finally, AECs are the predominant secretors of EC-derived stem cell factor (SCF) in the BM[26]. Conversely, SECs are part of more permeable sinusoidal vessels, resulting in high plasma penetration and exposure of perivascular HSCs to higher levels of ROS[22,27]. They express high levels of C-X-C motif chemokine ligand 12 (CXCL12), required for stem cell homing[28]. Altogether, these data show that AECs are thought to support more primitive, quiescent HSCs, whereas SECs support HSC proliferation and mobilization[29].

Increasing evidence indicates that ECs play an important role in hematopoietic recovery after HCT, by production of several hematopoietic stem and progenitor cell (HSPC)-supporting molecules. In mice, engraftment of transplanted HSPCs after either 5-fluorouracil (5-FU) or irradiation depends on recovery of SECs, which is mediated through activation of vascular endothelial growth factor receptor 2 (VEGFR2) signaling[30]. Inhibition of this signaling, through conditional deletion of VEGFR2, results in disorganized regeneration of SECs, delayed hematopoietic recovery and persistent life-threatening pancytopenia[30]. Furthermore, EC-specific expression of Tie2[31], Jagged-1[32] and Jagged-2[33] have all been shown to support hematopoietic regeneration after myeloablative injury, by promoting regeneration of the vascular niche (Tie2) or by activating Notch signaling in HSPCs (Jagged-1 and Jagged-2). Finally, a subtype of capillary ECs expressing ***Apelin*** (Apln$^+$ ECs), increases substantially after irradiation and is critical for post-transplant hematopoietic recovery in mice[34]. Interestingly, elimination of HSPCs by diphtheria toxin phenocopied the vascular changes observed after irradiation or 5-FU, indicating that HSCs actively maintain their niche, and vice versa[34].

In humans, recent studies identified a subset of BM ECs, CD105 (*endoglin*)-expressing ECs, which are nearly absent during homeostatic hematopoiesis but are enriched in fetal BM and during regeneration upon chemotherapeutic injury[24]. These ECs express high levels of interleukin-33 (IL-33), which promotes the expansion of both hematopoietic precursor cells and other EC subsets *ex vivo*[24]. Interestingly, a subset of these cells, $CD105^{+}CD271^{+}$ ECs, co-express endothelial as well as stromal markers and have the potential to convert to stromal progenitor cells and their downstream progeny[24]. Upon subcutaneous implantation in mice, these human $CD105^{+}CD271^{+}$ EC-derived cells formed pellets consisting of human bone, cartilage, adipocytes and blood vessels which recruited hematopoietic cells, supporting their *in vivo* niche-regenerative capacity[23].

In the context of HCT, ECs are exposed to a variety of damaging stimuli. In mice, irradiation or administration of 5-FU causes loss of ECs in a dose-dependent manner[30]. In human patients, conditioning with high-dose cyclophosphamide or busulfan is associated with increased risk of EC-related disorders, such as veno-occlusive disease/sinusoidal obstruction syndrome, thrombotic microangiopathy, capillary leak syndrome and idiopathic pneumonia syndrome[35]. Furthermore, bacterial endotoxins, inflammatory cytokines and calcineurin inhibitors have all been associated with EC injury[35], which could in turn impair hematopoietic recovery.

ECs may provide an opportunity to promote hematopoietic recovery after HCT, by co-infusion of healthy ECs with the stem cell product, or by protecting these cells from conditioning-induced damage. In mice, co-infusion of ECs together with hematopoietic cells improves HSC repopulating activity, engraftment and survival after irradiation, compared to infusion of hematopoietic cells only[36]. The beneficial effect of EC co-infusion on hematopoietic recovery is even more prominent when the ECs are pre-treated with the Wnt-antagonist Dickkopf1 (Dkk1), which induces secretion of several proteins known to promote hematopoietic regeneration, including granulocyte colony-stimulating factor (G-CSF) and VEGF[37]. Strategies to protect recipient ECs from chemotherapy-induced damage include administration of pigment endothelial derived factor (PEDF)[38], defibrotide[39] and N-acetyl-L-cysteine (NAC)[40]. Although PEDF and defibrotide have been shown to improve hematopoietic recovery in mice, their effects in humans are still unknown. Prophylactic oral NAC treatment was shown to be safe and effective in preventing poor hematopoietic reconstitution in human HCT-recipients, suggested to be a result of improved BM EC function

[40]. A Phase III, open-labeled, randomized clinical trials is currently recruiting to further investigate NAC for prevention of poor hematopoietic reconstitution in patients receiving an HCT (Trial no. NCT03967665).

Mesenchymal stromal cells

Mesenchymal stromal cells (MSCs) are a rare (~0.001%–0.01%) component of the BM niche. MSCs were first described in 1968 as a population of adherent cells of the BM, which exhibited a fibroblast-like morphology and which can differentiate *in vitro* into bone, cartilage, adipose tissue, tendon and muscle[41]. BM MSCs co-localize closely with HSCs and regulate HSC homeostasis through the production of soluble factors, including CXCL12, angiopoietin and SCF, which are key factors for HSC maintenance. In recent years, advances in flow cytometry and cell-tracing methods have identified multiple MSC subsets, with distinct impact on HSC behavior. $CD271^+$ and $CD271^+/CD146^{-/low}$ MSCs are bone-lining cells that support long-term, quiescent HSCs in areas with low oxygen tension. In contrast, $CD271^+/CD146^+$ MSCs are located in the perivascular region where they support more proliferative HSCs[42]. In mice, a specific type of perivascular MSCs, $Nestin^+/NG2^+$ MSCs, produces high levels of CXCL12 and angiopoietin[20,43]. Depletion of these cells using Nestin-Cre results in loss of HSCs, supporting the HSC-supporting role of these cells in the murine BM niche.

MSCs are extensively studied in the context of HCT. In the majority of HCT recipients, MSCs remain of recipient origin, indicating that these cells are not fully eradicated by myeloablative conditioning[44,45]. The mechanisms that allow MSCs to survive pre-HCT conditioning regimens that are lethal to hematopoietic cells remain incompletely understood, and may involve more efficient recognition of DNA damage, double strand break repair and evasion of apoptosis[46,47]. Conversely, one might hypothesize that MSCs can simply tolerate a higher mutational load than hematopoietic cells, for instance, by expressing translesion synthesis polymerases[48]. Although recipient MSCs may remain relatively viable after conditioning, they do accumulate damage[46]. For example, *in vitro* irradiation of human MSCs results in accumulation of DNA double-strand breaks[49], altered gene expression[50], skewed differentiation towards osteogenesis[51] and induction of senescence[52]. Interestingly, reports have shown that in recipients transplanted with BM and PB grafts, part of the MSC pool after HCT was of donor-origin[53,54]. Therefore, it will be of interest to investigate how donor-recipient MSC chimerism and conditioning-induced damage relate to post-HCT hematopoietic function.

Because of their regenerative and immune-regulatory properties, MSCs are used as a clinical therapy for a variety of degenerative and inflammatory diseases, including articular cartilage defects, cardiac diseases, inflammatory bowel disease and severe COVID-19[55]. In the context of HCT, MSCs have been used to enhance HSC engraftment and to treat steroid-resistant aGvHD. The use of MSCs for aGvHD is beyond the scope of this review and has been reviewed elsewhere[56]. In phase I/II trials in human allo-HCT recipients, co-infusion of MSCs together with hematopoietic cells was safe and resulted in prompt engraftment in 144 out of 146 recipients[56], compared to 5-10% risk of graft failure in historic controls. Whether this apparent improvement is due to niche-restoring or immunosuppressive effects remains to be defined. Thus far, no comparative phase III studies have studied the role of MSC infusions in the prevention or treatment of non-engraftment after HCT. The feasibility of such studies is hampered by the rarity of graft rejection, the heterogeneity of the patient group and of the MSC cell product, thus requiring large numbers of patients. To facilitate such studies, it will be of interest to investigate the niche prior to HCT, to identify potential biomarkers of increased niche damage in HCT recipients, who are most likely to benefit from niche-correcting strategies.

Osteolineage cells

Osteolineage cells are a heterogeneous pool of bone-forming cells of various developmental stages, including pre-osteoblasts, osteoblasts and terminally differentiated osteocytes[57]. Osteolineage cells were among the first niche cell types to be implicated in the regulation of HSCs[58]. Early mouse studies showed that long-term repopulating (LT-)HSCs co-localize closely with osteoblasts. Osteoblasts secrete several factors required for HSC maintenance, such as CXCL12 [59], SCF[59], angiopoietin[60], thrombopoietin[61] and osteopontin (OPN)[59]. Finally, the number of osteoblasts in the niche is closely correlated with the number of HSCs[59,62], and conditional ablation of osteoblasts results in loss of lymphoid, erythroid and myeloid hematopoietic progenitor cells from the BM[63].

However, more recently, the role of osteolineage cells in HSC regulation has been subject of debate. For instance, whereas osteoblasts produce HSC-supporting molecules, they may not be the predominant source of these factors. Hepatocytes, and not BM cells, are likely the major source of thrombopoietin[64] and HSCs and stromal cells are the main producers of BM angiopoietin[65]. In addition, selective deletion of *CXCL12* or *SCF* from murine osteoblasts has little effect on HSCs[28,66]. Furthermore, recent 3D imaging studies in mice have shown that the majority of endogenous HSCs lie adjacent to BM blood vessels, in close

association with endothelial and mesenchymal cells, and that only a minority of HSCs is localized in direct contact with BM osteoblasts[19,67]. In summary, these studies suggest that osteolineage cells may be less important for HSC maintenance during homeostatic hematopoiesis than previously thought. Notably, osteolineage cells have been shown to regulate more committed hematopoietic progenitor cells in mice[59,65,66], and their potential role during hematopoietic regeneration, in mice as well as in humans, remains to be defined.

Conditioning-induced damage to osteolineage cells is thought to underlie bone-related complications after allo-HCT, including bone loss, osteopenia, osteoporosis and avascular necrosis of bone[68]. *In vitro* chemotherapeutic treatment of murine[69] as well as human[70] osteoblasts with VP16 or melphalan resulted in decreased production of CXCL12 and reduced capacity to support immature B progenitor cells and CD34+ BM cells[70]. Similarly, irradiation induces several functional defects in osteoblasts, such as decreased production of extracellular matrix components[71], impaired proliferation[71] and induction of apoptosis[72]. Notably, in addition to pre-HCT conditioning, various other HCT-related exposures may compromise osteoblast numbers and/or function after HCT, including corticosteroids[73], calcineurin inhibitors[68], nutritional deficiencies and G-CSF[74].

Several studies have attempted to prevent and/or restore conditioning-induced damage to osteolineage cells, to prevent bone complications after HCT and/or to accelerate hematopoietic recovery. Strategies for prevention and treatment of bone loss are excellently reviewed by McCune et al[68]. In mice, parathyroid hormone (PTH) injection increases the number of osteoblasts and HSCs in the BM, and improves post-HCT survival[62]. However, a subsequent phase II study in human HCT recipients was halted early because of excessive treatment-related mortality[75,76]. In the 13 evaluable patients, no beneficial effect of PTH on hematopoietic engraftment was observed[75,76], again suggesting that the impact of osteolineage cells on hematopoietic recovery may be less evident than previously thought.

Adipocytes

Bone marrow adipocytes (BMAs) differentiate from MSCs and comprise a heterogeneous population of cells. Although BMAs were initially considered simple "fillers" of marrow space, increasing evidence indicates that they actively contribute to hematopoiesis[77,78]. BMAs produce adiponectin, which stimulates

HSC proliferation *in vitro*[79]. During ageing, the number of adipocytes in the BM niche increases progressively, gradually replacing sites with hematopoietic activity[59,80]. Furthermore, in mice, adipocyte content differs between different bones and is negatively correlated with HSPC content[80].

In mice[81] as well as in humans[82], chemotherapy and irradiation are associated with increased BM adipocyte content, potentially contributing to (transient) hematopoietic aplasia. Depletion of BM adipocytes, either by genetic engineering (fat-free A-ZIP/F1 mice) or by treatment with the PPARγ inhibitor Bisphenol-A-DiGlycidyl-Ether, resulted in accelerated hematopoietic recovery after irradiation[80,83]. Conversely, BMAs have also been reported to promote hematopoietic regeneration. For instance, adiponectin-null mice showed delayed hematopoietic recovery upon myeloablative injury compared to wild type mice[84]. Furthermore, murine BMAs produce SCF, and adipocyte-specific deletion of SCF inhibited hematopoietic regeneration after irradiation or chemotherapy, resulting in increased transplant-related mortality[77]. Interestingly, treatment of murine HCT recipients with simvastatin, a drug already used in the treatment of hypercholesterolemia, prevents radiotherapy-induced BM adipogenesis and improves HSC engraftment[85]. Taken together, the impact of BMAs on steady-state hematopoiesis and hematopoietic regeneration remains controversial and requires future studies, particularly in humans.

Nerve fibers

BM nerves regulate the proliferation, differentiation, and migration between the BM and extramedullary sites of HSPCs, during homeostatic hematopoiesis and after HCT. Most studies have focused on the sympathetic nervous system (SNS) [20,85], although more recently, a role for the parasympathetic nervous system was also proposed[87]. Sympathetic nerve fibers release noradrenalin, which facilitates HSPC egression from the BM towards extramedullary sites[21]. In fact, circadian changes in the balance between sympathetic and parasympathetic signaling are thought to underlie the daily oscillations in HSC proliferation and migration, as reviewed by Mendez-Ferrer et al (2009)[88]. The interaction between the SNS and HSCs is (at least in part) mediated via niche cells, as binding of noradrenalin to the β_3 adrenergic receptor expressed by stromal cells results in downregulation of CXCL12, the key niche retention chemokine[89]. In addition, noradrenalin-mediated activation of the β_3-adrenergic receptor on HSPCs promotes HPSC mobility and proliferation[90]. The importance of sympathetic nerve signaling for HSC maintenance is exemplified by recent murine studies, demonstrating that loss of sympathetic nerves or β_3 adrenergic signaling in the

BM results in premature HSC ageing, which can be rescued by supplementation of sympathomimetic agents[91].

Chemotherapy and/or irradiation can be particularly neurotoxic, inducing transient or persistent sympathetic neuropathy which may contribute to hematopoietic dysfunction.

In mice, chemotherapy with cisplatin or 5-FU is associated with decreased numbers of SNS fibers in the BM[10]. In humans, many cancer survivors suffer from radiation-induced neuropathy[92]. Similarly, several chemotherapeutic drugs (e.g. vinca alkaloids, taxanes, platinum-based agents) and calcineurin-inhibitors commonly induce severe peripheral neuropathy[93,94].

Whether and how chemo- or radiotherapy-induced neuropathy impacts on post-transplant hematopoietic regeneration remains incompletely understood. In mice, cisplatin-induced sensory neuropathy is associated with impaired bone marrow regeneration and decreased survival after HCT[10]. In these mice, selective depletion of adrenergic innervation in the BM by 6-hydroxydopamine resulted in prolonged BM aplasia, both after chemotherapeutic myeloablation as well as after irradiation[10]. This effect was specific to neurons, because protection from chemotherapy-induced nerve damage by deletion of *Trp53* in sympathetic neurons, or by administration of neurotrophic compounds, could restore hematopoietic recovery[10]. Furthermore, administration of hematopoietic growth factors, such as G-CSF and granulocyte-macrophage colony-stimulating factor (GM-CSF) is associated with increased expression of neuronal receptors on HSPCs, enhancing their proliferation and repopulation capacity[90]. Importantly, although sympathetic neurons are important regulators of HSCs, they also impact on the behavior of other niche cell types, for example MSCs, thereby indirectly influencing hematopoietic cells[8689].

Beyond the SNS, the nociceptive nervous system has also been shown to impacts on HSC homing and migration[95]. Treatment of mice with calcitonin gene-related peptide (CGRP), the main nociceptive neurotransmitter, substantially increased G-CSF induced HSC mobilization into the peripheral blood, at the expense of BM HSC content[95]. CGRP interacts directly with HSCs, via receptor activity modifying protein 1 (RAMP1) and the calcitonin receptor-like receptor (CALCRL), increasing intracellular cAMP levels which facilitate HSC mobilization. Intriguingly, mice fed capsaicin-containing food, a known nociceptive activator, also displayed significantly enhanced HSC mobilization. As HCT is associated with many painful

stimuli and analgesic medications, it will be of interest to investigate the impact of nociceptive signaling on hematopoietic recovery in this context.

Conclusions

Recent technological advances have allowed deconstruction of the BM niche and provide insight into the mechanisms by which the niche is affected by HCT conditioning and how it regulates HSC behavior, during homeostatic hematopoiesis and hematopoietic regeneration. HCT is associated with multiple changes in the BM niche, including dysfunction of ECs and neurons, accumulation of DNA damage in MSCs, reduced numbers osteoprogenitor cells and increased numbers of adipocytes, which may collectively impair hematopoietic reconstitution (Table 1). To improve HCT outcome, several niche-directed strategies have been explored, including antibody-based conditioning[96,97], infusion of extracellular vesicles derived from BM-MSCs[98,99], co-infusion of HSCs with autologous or allogeneic MSCs or ECs[36,100], inhibition of adipogenesis by simvastatin treatment[85], or the use of endothelial[40,3839] and neuroprotective compounds[10] (Table 2).

Importantly, as interplay between several functionally intact niche cell types is likely required for adequate HSC support, combination therapies may be required. Future studies are needed to compare the impact of different pre-HCT conditioning regimens on the BM niche in mice as well as in humans, to identify the BM niche cell types most susceptible to conditioning-induced damage, to assess the impact of this damage on HSC engraftment and long-term function, and to select the most appropriate treatment.

List of abbreviations

5-FU: 5-fluorouracil
AEC: Arteriolar endothelial cell
BM: Bone marrow
BMA: Bone marrow adipocyte
CGRP: Calcitonin gene-related peptide
CXCL12: C-X-C motif chemokine ligand 12
EC: Endothelial cell
G-CSF: Granulocyte-colony stimulating factor

GM-CSF: Granulocyte-macrophage colony stimulating factor
HCT: Hematopoietic cell transplantation
HSC: Hematopoietic stem cell
MSC: Mesenchymal stromal cell
RAMP1: Receptor activity modifying protein 1
ROS: Reactive oxygen species
SCF: Stem cell factor
SEC: Sinusoidal endothelial cell
SNS: Sympathetic nervous system
VEGFR2: Vascular endothelial growth factor receptor 2

Declarations

Ethics approval and consent to participate:
Not applicable.

Consent for publication:
Not applicable.

Availability of data and materials:
Not applicable.

Competing interests:
The authors declare no competing interests.

Funding
M.E.B. receives financial support by research grants of the Dutch National Research
Council (VI.Veni.202.021), by a Leukemia Fellowship Grant of the European Society for Blood
and Marrow Transplantation and by a John Hansen Research Grant of the DKMS. These funding bodies had no role in the design of the study, nor in the collection, analysis, and interpretation of the data, nor in writing the manuscript.

Author's contributions
F.P., L.D., A.P. and M.E.B. conceptualized the manuscript. F.P., L.D., and M.E.B wrote the manuscript. R.v.B and S.N. provided essential comments and feedback. All authors read and approved the final manuscript.

Acknowledgements:
Not applicable.

References

1 Passweg JR, Baldomero H, Chabannon C, Basak GW, de la Cámara R, Corbacioglu S *et al.* Hematopoietic cell transplantation and cellular therapy survey of the EBMT: monitoring of activities and trends over 30 years. *Bone Marrow Transplantation* 2021. doi:10.1038/s41409-021-01227-8.

2 Phelan R, Arora M, Chen M. Current use and outcome of hematopoietic stem cell transplantation: CIBMTR US summary slides, 2020. .

3 Olsson R, Remberger M, Schaffer M, Berggren DM, Svahn B-M, Mattsson J *et al.* Graft failure in the modern era of allogeneic hematopoietic SCT. *Bone Marrow Transplantation* 2013; **48**: 537–543.

4 Sun Y-Q, He G-L, Chang Y-J, Xu L-P, Zhang X-H, Han W *et al.* The incidence, risk factors, and outcomes of primary poor graft function after unmanipulated haploidentical stem cell transplantation. *Annals of Hematology* 2015; **94**: 1699–1705.

5 de Koning C, Langenhorst J, van Kesteren C, Lindemans CA, Huitema ADR, Nierkens S *et al.* Innate Immune Recovery Predicts CD4+ T Cell Reconstitution after Hematopoietic Cell Transplantation. *Biology of Blood and Marrow Transplantation* 2019; **25**: 819–826.

6 Velardi E, Clave E, Arruda LCM, Benini F, Locatelli F, Toubert A. The role of the thymus in allogeneic bone marrow transplantation and the recovery of the peripheral T-cell compartment. *Seminars in Immunopathology* 2021; **43**: 101–117.

7 de Koning C, Nierkens S, Boelens JJ. Strategies before, during, and after hematopoietic cell transplantation to improve T-cell immune reconstitution. *Blood* 2016; **128**: 2607–2615.

8 Schofield R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood Cells* 1978; **4**: 7—25.

9 Mendelson A, Frenette PS. Hematopoietic stem cell niche maintenance during homeostasis and regeneration. *Nature Medicine* 2014; **20**: 833–846.

10 Lucas D, Scheiermann C, Chow A, Kunisaki Y, Bruns I, Barrick C *et al.* Chemotherapy-induced bone marrow nerve injury impairs hematopoietic regeneration. *Nature Medicine* 2013; **19**: 695–703.

11 Cao X, Wu X, Frassica D, Yu B, Pang L, Xian L *et al.* Irradiation induces bone injury by damaging bone marrow microenvironment for stem cells. *Proc Natl Acad Sci U S A* 2011; **108**: 1609–1614.

12 Pronk E, Raaijmakers MHGP. The mesenchymal niche in MDS. *Blood* 2019; **133**: 1031–1038.

13 Kong Y, Chang Y-J, Wang Y-Z, Chen Y-H, Han W, Wang Y *et al.* Association of an Impaired Bone Marrow Microenvironment with Secondary Poor Graft Function after Allogeneic Hematopoietic Stem Cell Transplantation. *Biology of Blood and Marrow Transplantation* 2013; **19**: 1465–1473.

14 Tikhonova AN, Dolgalev I, Hu H, Sivaraj KK, Hoxha E, Cuesta-Domínguez Á *et al.* The bone marrow microenvironment at single-cell resolution. *Nature* 2019; **569**: 222–228.

15 Baccin C, Al-Sabah J, Velten L, Helbling PM, Grünschläger F, Hernández-Malmierca P *et al.* Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nature Cell Biology* 2020; **22**: 38–48.

16 Spencer JA, Ferraro F, Roussakis E, Klein A, Wu J, Runnels JM *et al.* Direct measurement of local oxygen concentration in the bone marrow of live animals. *Nature* 2014; **508**: 269–273.

17 Simsek T, Kocabas F, Zheng J, DeBerardinis RJ, Mahmoud AI, Olson EN *et al.* The Distinct Metabolic Profile of Hematopoietic Stem Cells Reflects Their Location in a Hypoxic Niche. *Cell Stem Cell* 2010; **7**: 380–390.
18 Takubo K, Goda N, Yamada W, Iriuchishima H, Ikeda E, Kubota Y *et al.* Regulation of the HIF-1α Level Is Essential for Hematopoietic Stem Cells. *Cell Stem Cell* 2010; **7**: 391–402.
19 Kunisaki Y, Bruns I, Scheiermann C, Ahmed J, Pinho S, Zhang D *et al.* Arteriolar niches maintain haematopoietic stem cell quiescence. *Nature* 2013; **502**: 637–643.
20 Méndez-Ferrer S, Michurina T V., Ferraro F, Mazloom AR, MacArthur BD, Lira SA *et al.* Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. *Nature* 2010; **466**: 829–834.
21 Katayama Y, Battista M, Kao WM, Hidalgo A, Peired AJ, Thomas SA *et al.* Signals from the sympathetic nervous system regulate hematopoietic stem cell egress from bone marrow. *Cell* 2006; **124**: 407–421.
22 Itkin T, Gur-Cohen S, Spencer JA, Schajnovitz A, Ramasamy SK, Kusumbe AP *et al.* Distinct bone marrow blood vessels differentially regulate haematopoiesis. *Nature* 2016; **532**: 323–328.
23 Kenswil KJG, Pisterzi P, Sánchez-Duffhues G, van Dijk C, Lolli A, Knuth C *et al.* Endothelium-derived stromal cells contribute to hematopoietic bone marrow niche formation. *Cell Stem Cell* 2021; **28**: 653–670.
24 Kenswil KJG, Jaramillo AC, Ping Z, Chen S, Hoogenboezem RM, Mylona MA *et al.* Characterization of Endothelial Cells Associated with Hematopoietic Niche Formation in Humans Identifies IL-33 As an Anabolic Factor. *Cell Reports* 2018; **22**: 666–678.
25 Renders S, Svendsen AF, Panten J, Rama N, Maryanovich M, Sommerkamp P *et al.* Niche derived netrin-1 regulates hematopoietic stem cell dormancy via its receptor neogenin-1. *Nature Communications* 2021; **12**: 1–15.
26 Xu C, Gao X, Wei Q, Nakahara F, Zimmerman SE, Mar J *et al.* Stem cell factor is selectively secreted by arterial endothelial cells in bone marrow. *Nature Communications* 2018; **9**: 1–13.
27 Ludin A, Gur-Cohen S, Golan K, Kaufmann KB, Itkin T, Medaglia C *et al.* Reactive oxygen species regulate hematopoietic stem cell self-renewal, migration and development, as well as their bone marrow microenvironment. *Antioxidants and Redox Signaling* 2014; **21**: 1605–1619.
28 Ding L, Saunders TL, Enikolopov G, Morrison SJ. Endothelial and perivascular cells maintain haematopoietic stem cells. *Nature* 2012; **481**: 457–462.
29 Itkin T, Gur-Cohen S, Spencer JA, Schajnovitz A, Ramasamy SK, Kusumbe AP *et al.* Distinct bone marrow blood vessels differentially regulate haematopoiesis. *Nature* 2016; **532**: 323–328.
30 Hooper AT, Butler JM, Nolan DJ, Kranz A, Iida K, Kobayashi M *et al.* Engraftment and Reconstitution of Hematopoiesis Is Dependent on VEGFR2-Mediated Regeneration of Sinusoidal Endothelial Cells. *Cell Stem Cell* 2009; **4**: 263–274.
31 Kopp H-G, Avecilla ST, Hooper AT, Shmelkov S V., Ramos CA, Zhang F *et al.* Tie2 activation contributes to hemangiogenic regeneration after myelosuppression. *Blood* 2005; **106**: 505–513.
32 Poulos MG, Guo P, Kofler NM, Pinho S, Gutkin MC, Tikhonova A *et al.* Endothelial Jagged-1 Is Necessary for Homeostatic and Regenerative Hematopoiesis. *Cell Reports* 2013; **4**: 1022–1034.

33 Guo P, Poulos MG, Palikuqi B, Badwe CR, Lis R, Kunar B *et al.* Endothelial jagged-2 sustains hematopoietic stem and progenitor reconstitution after myelosuppression. *Journal of Clinical Investigation* 2017; **127**: 4242–4256.

34 Chen Q, Liu Y, Jeong HW, Stehling M, Dinh VV, Zhou B *et al.* Apelin+ Endothelial Niche Cells Control Hematopoiesis and Mediate Vascular Regeneration after Myeloablative Injury. *Cell Stem Cell* 2019; **25**: 768-783.e6.

35 Hildebrandt GC, Chao N. Endothelial cell function and endothelial-related disorders following haematopoietic cell transplantation. *British Journal of Haematology* 2020; **190**: 508–519.

36 Poulos MG, Ramalingam P, Gutkin MC, Llanos P, Gilleran K, Rabbany SY *et al.* Endothelial transplantation rejuvenates aged hematopoietic stem cell function. *Journal of Clinical Investigation* 2017; **127**: 4163–4178.

37 Kim MM, Schlussel L, Zhao L, Himburg HA. Dickkopf-1 Treatment Stimulates Hematopoietic Regenerative Function in Infused Endothelial Progenitor Cells. *Radiation Research* 2019; **192**: 53–62.

38 Ju W, Lu W, Ding L, Bao Y, Hong F, Chen Y *et al.* PEDF promotes the repair of bone marrow endothelial cell injury and accelerates hematopoietic reconstruction after bone marrow transplantation. *Journal of Biomedical Science* 2020; **27**: 1–13.

39 Eissner G, Multhoff G, Gerbitz A, Kirchner S, Bauer S, Haffner S *et al.* Fludarabine induces apoptosis, activation, and allogenicity in human endothelial and epithelial cells: Protective effect of defibrotide. *Blood* 2002; **100**: 334–340.

40 Kong Y, Wang Y, Zhang YY, Shi MM, Mo XD, Sun YQ *et al.* Prophylactic oral NAC reduced poor hematopoietic reconstitution by improving endothelial cells after haploidentical transplantation. *Blood Advances* 2019; **3**: 1303–1317.

41 Viswanathan S, Shi Y, Galipeau J, Krampera M, Leblanc K, Martin I *et al.* Mesenchymal stem versus stromal cells: International Society for Cell & Gene Therapy (ISCT®) Mesenchymal Stromal Cell committee position statement on nomenclature. *Cytotherapy* 2019; **21**: 1019–1024.

42 Tormin A, Li O, Brune JC, Walsh S, Schütz B, Ehinger M *et al.* CD146 expression on primary nonhematopoietic bone marrow stem cells is correlated with in situ localization. *Blood* 2011; **117**: 5067–5077.

43 Asada N, Kunisaki Y, Pierce H, Wang Z, Fernandez NF, Birbrair A *et al.* Differential cytokine contributions of perivascular haematopoietic stem cell niches. *Nature Cell Biology* 2017; **19**: 214–223.

44 García-Castro J, Balas A, Ramírez M, Pérez-Martínez A, Madero L, González-Vicent M *et al.* Mesenchymal stem cells are of recipient origin in pediatric transplantations using umbilical cord blood, peripheral blood, or bone marrow. *Journal of Pediatric Hematology/Oncology* 2007; **29**: 388–392.

45 Rieger K, Marinets O, Fietz T, Körper S, Sommer D, Mücke C *et al.* Mesenchymal stem cells remain of host origin even a long time after allogeneic peripheral blood stem cell or bone marrow transplantation. *Experimental Hematology* 2005; **33**: 605–611.

46 Nicolay NH, Perez RL, Saffrich R, Huber PE. Radio-resistant mesenchymal stem cells: Mechanisms of resistance and potential implications for the clinic. *Oncotarget* 2015; **6**: 19366–19380.

47 Alessio N, Del Gaudio S, Capasso S, Di Bernardo G, Cappabianca S, Cipollaro M *et al.* Low dose radiation induced senescence of human mesenchymal stromal cells and impaired the autophagy process. *Oncotarget* 2015; **6**: 8155–8166.

48 Lange SS, Takata K, Wood RD. DNA polymerases and cancer. *Nature Reviews Cancer* 2011; **11**: 96–110.

49 Sugrue T, Lowndes NF, Ceredig R. Mesenchymal stromal cells: radio-resistant members of the bone marrow. *Immunology & Cell Biology* 2013; **91**: 5–11.

50 Preciado S, Muntión S, Rico A, Pérez-Romasanta LA, Ramos TL, Ortega R *et al.* Mesenchymal Stromal Cell Irradiation Interferes with the Adipogenic/Osteogenic Differentiation Balance and Improves Their Hematopoietic-Supporting Ability. *Biology of Blood and Marrow Transplantation* 2018; **24**: 443–451.

51 Preciado S, Muntión S, Rico A, Pérez-Romasanta LA, Ramos TL, Ortega R *et al.* Mesenchymal Stromal Cell Irradiation Interferes with the Adipogenic/Osteogenic Differentiation Balance and Improves Their Hematopoietic-Supporting Ability. *Biology of Blood and Marrow Transplantation* 2018; **24**: 443–451.

52 Nicolay NH, Rühle A, Perez RL, Trinh T, Sisombath S, Weber KJ *et al.* Mesenchymal stem cells exhibit resistance to topoisomerase inhibition. *Cancer Letters* 2016; **374**: 75–84.

53 Poloni A, Leoni P, Buscemi L, Balducci F, Pasquini R, Masia MC *et al.* Engraftment capacity of mesenchymal cells following hematopoietic stem cell transplantation in patients receiving reduced-intensity conditioning regimen. *Leukemia* 2006; **20**: 329–335.

54 Villaron EM, Almeida J, Lopez-Holgado N, Alcoceba M, Sanchez-Abarca LI, Sanchez-Guijo FM *et al.* Mesenchymal stem cells are present in peripheral blood and can engraft after allogeneic hematopoietic stem cell transplantation. *Haematologica* 2004; **89**: 1421–7.

55 Galipeau J, Sensébé L. Mesenchymal Stromal Cells: Clinical Challenges and Therapeutic Opportunities. *Cell Stem Cell* 2018; **22**: 824–833.

56 Andre J. Burnham, Lisa P. Daley-Bauer and EMH. Mesenchymal stromal cells in hematopoietic stem cell transplantation. *Blood Advances* 2020; **4**: 5877–5887.

57 Asada N, Katayama Y. Regulation of hematopoiesis in endosteal microenvironments. International Journal of Hematology. 2014; **99**: 679–684.

58 Lo Celso C, Fleming HE, Wu JW, Zhao CX, Miake-Lye S, Fujisaki J *et al.* Live-animal tracking of individual haematopoietic stem/progenitor cells in their niche. *Nature* 2009; **457**: 92–96.

59 Pinho S, Frenette PS. Haematopoietic stem cell activity and interactions with the niche. *Nature Reviews Molecular Cell Biology* 2019; **20**: 303–320.

60 Arai F, Hirao A, Ohmura M, Sato H, Matsuoka S, Takubo K *et al.* Tie2/Angiopoietin-1 Signaling Regulates Hematopoietic Stem Cell Quiescence in the Bone Marrow Niche. *Cell* 2004; **118**: 149–161.

61 Yoshihara H, Arai F, Hosokawa K, Hagiwara T, Takubo K, Nakamura Y *et al.* Thrombopoietin/MPL Signaling Regulates Hematopoietic Stem Cell Quiescence and Interaction with the Osteoblastic Niche. *Cell Stem Cell* 2007; **1**: 685–697.

62 Calvi LM, Adams GB, Weibrecht KW, Weber JM, Olson DP, Knight MC *et al.* Osteoblastic cells regulate the haematopoietic stem cell niche. *Nature* 2003; **425**: 841–846.

63 Visnjic D, Kalajzic Z, Rowe DW, Katavic V, Lorenzo J, Aguila HL. Hematopoiesis is severely altered in mice with an induced osteoblast deficiency. *Blood* 2004; **103**: 3258–3264.

64 Decker M, Leslie J, Liu Q, Ding L. Hepatic thrombopoietin is required for bone marrow hematopoietic stem cell maintenance. *Science (1979)* 2018; **360**: 106–110.

65 Ding L, Morrison SJ. Haematopoietic stem cells and early lymphoid progenitors occupy distinct bone marrow niches. *Nature* 2013; **495**: 231–235.

66 Greenbaum A, Hsu Y-MS, Day RB, Schuettpelz LG, Christopher MJ, Borgerding JN *et al.* CXCL12 in early mesenchymal progenitors is required for haematopoietic stem-cell maintenance. *Nature* 2013; **495**: 227–230.

67 Nombela-Arrieta C, Pivarnik G, Winkel B, Canty KJ, Harley B, Mahoney JE *et al.* Quantitative imaging of haematopoietic stem and progenitor cell localization and hypoxic status in the bone marrow microenvironment. *Nature Cell Biology* 2013; **15**: 533–543.

68 McClune B, Majhail NS, Flowers MED. Bone Loss and Avascular Necrosis of Bone After Hematopoietic Cell Transplantation. *Seminars in Hematology* 2012; **49**: 59–65.

69 Gencheva M, Hare I, Kurian S, Fortney J, Piktel D, Wysolmerski R *et al.* Bone marrow osteoblast vulnerability to chemotherapy. *European Journal of Haematology* 2013; **90**: 469–478.

70 Rellick SL, O'Leary H, Piktel D, Walton C, Fortney JE, Akers SM *et al.* Bone marrow osteoblast damage by chemotherapeutic agents. *PLoS ONE* 2012; **7**: e30758.

71 Lau P, Baumstark-Khan C, Hellweg CE, Reitz G. X-irradiation-induced cell cycle delay and DNA double-strand breaks in the murine osteoblastic cell line OCT-1. *Radiation and Environmental Biophysics* 2010; **49**: 271–280.

72 Szymczyk KH, Shapiro IM, Adams CS. Ionizing radiation sensitizes bone cells to apoptosis. *Bone* 2004; **34**. doi:10.1016/j.bone.2003.09.003.

73 McDonough AK, Curtis JR, Saag KG. The epidemiology of glucocorticoid-associated adverse events. *Current Opinion in Rheumatology* 2008; **20**. doi:10.1097/BOR.0b013e3282f51031.

74 Winkler IG, Pettit AR, Raggatt LJ, Jacobsen RN, Forristal CE, Barbier V *et al.* Hematopoietic stem cell mobilizing agents G-CSF, cyclophosphamide or AMD3100 have distinct mechanisms of action on bone marrow HSC niches and bone formation. *Leukemia* 2012; **26**. doi:10.1038/leu.2012.17.

75 Ballen K, Mendizabal AM, Cutler C, Politikos I, Jamieson K, Shpall EJ *et al.* Phase II Trial of Parathyroid Hormone after Double Umbilical Cord Blood Transplantation. *Biology of Blood and Marrow Transplantation* 2012; **18**. doi:10.1016/j.bbmt.2012.06.016.

76 Li S, Zou D, Li C, Meng H, Sui W, Feng S *et al.* Targeting stem cell niche can protect hematopoietic stem cells from chemotherapy and G-CSF treatment. *Stem Cell Research & Therapy* 2015; **6**: 1–10.

77 Zhou BO, Yu H, Yue R, Zhao Z, Rios JJ, Naveiras O *et al.* Bone marrow adipocytes promote the regeneration of stem cells and haematopoiesis by secreting SCF. *Nature Cell Biology* 2017; **19**: 891–903.

78 Pinho S, Marchand T, Yang E, Wei Q, Nerlov C, Frenette PS. Lineage-Biased Hematopoietic Stem Cells Are Regulated by Distinct Niches. *Developmental Cell* 2018; **44**: 634-641.e4.

79 DiMascio L, Voermans C, Uqoezwa M, Duncan A, Lu D, Wu J *et al.* Identification of Adiponectin as a Novel Hemopoietic Stem Cell Growth Factor. *The Journal of Immunology* 2007; **178**: 3511–3520.

80 Naveiras O, Nardi V, Wenzel PL, Hauschka P V., Fahey F, Daley GQ. Bone-marrow adipocytes as negative regulators of the haematopoietic microenvironment. *Nature* 2009; **460**: 259–263.

81 Cao X, Wu X, Frassica D, Yu B, Pang L, Xian L *et al.* Irradiation induces bone injury by damaging bone marrow microenvironment for stem cells. *Proc Natl Acad Sci U S A* 2011; **108**: 1609–1614.

82 Nguyen T-V, Melville A, Nath S, Story C, Howell S, Sutton R *et al.* Bone Marrow Recovery by Morphometry during Induction Chemotherapy for Acute Lymphoblastic Leukemia in Children. *PLOS ONE* 2015; **10**: e0126233.

83 Zhu R-JJ, Wu M-QQ, Li Z-JJ, Zhang Y, Liu K-YY. Hematopoietic recovery following chemotherapy is improved by BADGE-induced inhibition of adipogenesis. *International Journal of Hematology* 2013; **97**: 58–72.

84 Masamoto Y, Arai S, Sato T, Kubota N, Takamoto I, Kadowaki T *et al.* Adiponectin Enhances Quiescence Exit of Murine Hematopoietic Stem Cells and Hematopoietic Recovery Through mTORC1 Potentiation. *Stem Cells* 2017; **35**: 1835–1848.
85 Manmohan S Bajaj , Suprita S Ghode , Rohan S Kulkarni, Lalita S Limaye VPK. Simvastatin improves hematopoietic stem cell engraftment by preventing irradiation-induced marrow adipogenesis and radio-protecting the niche cells. *Haematologica* 2010; **100**: 323–327.
86 Hanoun M, Maryanovich M, Arnal-Estapé A, Frenette PS. Neural Regulation of Hematopoiesis, Inflammation, and Cancer. *Neuron* 2015; **86**: 360–373.
87 García-García A, Korn C, García-Fernández M, Domingues O, Villadiego J, Martín-Pérez D *et al.* Dual cholinergic signals regulate daily migration of hematopoietic stem cells and leukocytes. *Blood* 2019; **133**: 224–236.
88 Méndez-Ferrer S, Chow A, Merad M, Frenette PS. Circadian rhythms influence hematopoietic stem cells. *Current Opinion in Hematology* 2009; **16**: 235–242.
89 Méndez-Ferrer S, Lucas D, Battista M, Frenette PS. Haematopoietic stem cell release is regulated by circadian oscillations. *Nature* 2008; **452**: 442–447.
90 Spiegel A, Shivtiel S, Kalinkovich A, Ludin A, Netzer N, Goichberg P *et al.* Catecholaminergic neurotransmitters regulate migration and repopulation of immature human CD34+ cells through Wnt signaling. *Nature Immunology* 2007; **8**: 1123–1131.
91 Maryanovich M, Zahalka AH, Pierce H, Pinho S, Nakahara F, Asada N *et al.* Adrenergic nerve degeneration in bone marrow drives aging of the hematopoietic stem cell niche. *Nature Medicine* 2018; **24**: 782–791.
92 Delanian S, Lefaix J-L, Pradat P-F. Radiation-induced neuropathy in cancer survivors. *Radiotherapy and Oncology* 2012; **105**: 273–282.
93 Cavaletti G, Marmiroli P. Chemotherapy-induced peripheral neurotoxicity. *Nature Reviews Neurology* 2010; **6**: 500–507.
94 Zajaczkowską R, Kocot-Kępska M, Leppert W, Wrzosek A, Mika J, Wordliczek J. Mechanisms of chemotherapy-induced peripheral neuropathy. *International Journal of Molecular Sciences* 2019; **20**: 1–29.
95 Gao X, Zhang D, Xu C, Li H, Caron KM, Frenette PS. Nociceptive nerves regulate haematopoietic stem cell mobilization. *Nature* 2021; **589**: 591–596.
96 Palchaudhuri R, Saez B, Hoggatt J, Schajnovitz A, Sykes DB, Tate TA *et al.* Non-genotoxic conditioning for hematopoietic stem cell transplantation using a hematopoietic-cell-specific internalizing immunotoxin. *Nature Biotechnology* 2016; **34**: 738–745.
97 Czechowicz A, Palchaudhuri R, Scheck A, Hu Y, Hoggatt J, Saez B *et al.* Selective hematopoietic stem cell ablation using CD117-antibody-drug-conjugates enables safe and effective transplantation with immunity preservation. *Nature Communications 2019 10:1* 2019; **10**: 1–12.
98 Schoefinius JS, Brunswig-Spickenheier B, Speiseder T, Krebs S, Just U, Lange C. Mesenchymal Stromal Cell-Derived Extracellular Vesicles Provide Long-Term Survival After Total Body Irradiation Without Additional Hematopoietic Stem Cell Support. *Stem Cells* 2017; **35**: 2379–2389.
99 Xie H, Sun L, Zhang L, Liu T, Chen L, Zhao A *et al.* Mesenchymal stem cell-derived microvesicles support Ex vivo expansion of cord blood-derived CD34+ cells. *Stem Cells International* 2016; **2016**: 6493241.

100 Ball LM, Bernardo ME, Roelofs H, Lankester A, Cometa A, Egeler RM ***et al.*** Cotransplantation of ex vivo–expanded mesenchymal stem cells accelerates lymphocyte recovery and may reduce the risk of graft failure in haploidentical hematopoietic stem-cell transplantation. *Blood* 2007; **110**: 2764–2767.

Tables

Table 1: Causes and consequences of pre-HCT conditioning on the BM niche.

Cell type	Factors produced	Role in the normal HSC niche	Impact of HCT conditioning	Consequences on hematopoietic recovery	Model system	Ref.
Endothelial cells	CXCL12, SCF, Angiopoietin	Maintenance of quiescent HSCs (AECs), HSC migration and proliferation(SECs)	Loss of ECs in a IR dose-dependent manner; increased risk of EC-related disorders in human	Engraftment depends on recovery of SECs through activation of VEGFR2; signaling inhibition results in delayed hematopoietic recovery	Mice, human	30,35
Mesenchymal stromal cells	CXCL12, SCF, Angiopoietin	HSC homeostasis	MSCs are not fully eradicated but do accumulate DNA damage	Unclear, hematopoietic recovery might be delayed.	Human, in vitro	46,49,50,52
Osteolineage cells	CXCL12, SCF, Angiopoietin, Thrombopoietin, Osteopontin	Regulation of more committed hematopoietic progenitor cells; HSC maintenance by osteoblasts is subject of debate.	Bone-related complications; compromised osteoblast number and function.	Unclear. The supportive effect of osteolineage cells on HSCs may be less evident than previously thought.	Mice and human (in vitro)	68–76,78
Adipocytes	Adiponectin, Lectin, SCF	HSC proliferation	Increased BM adipocyte content	Controversial. Potentially contributing to (transient) hematopoietic aplasia. However, others report on their promotive role in hematopoietic regeneration.	Mice and human (in vitro)	79,84

Table 1: Causes and consequences of pre-HCT conditioning on the BM niche. (continued)

Cell type	Factors produced	Role in the normal HSC niche	Impact of HCT conditioning	Consequences on hematopoietic recovery	Model system	Ref.
Sympathetic neurons	Noradrenalin	HSPCs proliferation, differentiation, and migration	Transient or persistent sympathetic neuropathy; loss of sympathetic fibers	Impaired bone marrow regeneration	Mice	[10]
Nociceptive neurons	CGRP	HSC homing and migration	Unknown	Unknown	NA	NA

Table 2: Niche-directed therapeutic strategies

Niche cell type	Therapeutic strategy	Model system	Effect	Ref.
Endothelial cells	Co-infusion of ECs with hematopoietic stem cells; Administration of PEDF, defibrotide, and NAC.	Mice and humans (phase I/II and III)	Improved HSC repopulating activity, engraftment, and survival after irradiation; Protection of recipient ECs from chemotherapy-induced damage; Prophylactic oral NAC was safe and effective in preventing poor hematopoietic reconstitution by improving BM EC function in allo-HCT recipients; Effect will further be identified in phase III clinical trial.	[36], [37], [38], [39], [40] Trial no. NCT03967665 (currently recruiting)
Mesenchymal stromal cells	Co-infusion of MSCs with hematopoietic cells	Humans (phase I/II clinical trials)	Prompt engraftment of donor HSCs	Trials are reviewed in ref.[55], Trial no. NCT04247945 (currently recruiting)
Osteolineage cells	Parathyroid hormone (PTH) injection	Mice and Humans (halted at phase II)	Increases the number of osteoblasts and HSCs in the BM and improves post-HCT survival in mice. No beneficial effect on hematopoietic engraftment was observed in human HCT recipients.	[75]
Adipocytes	Simvastatin treatment	Mice	Prevents radiotherapy-induced BM adipogenesis and improves HSC engraftment	[85]
Sympathetic nervous system	Administration of hematopoietic growth factors, such as G-CSF and GM-CSF; Neuroprotection by administration of 4-methylcatechol	Mice	Increased expression of neuronal receptors on HSPCs, enhancing their proliferation and repopulation capacity; Accelerates BM regeneration	[90]
Nociceptive neurons	Unknown	NA	NA	NA

Figures

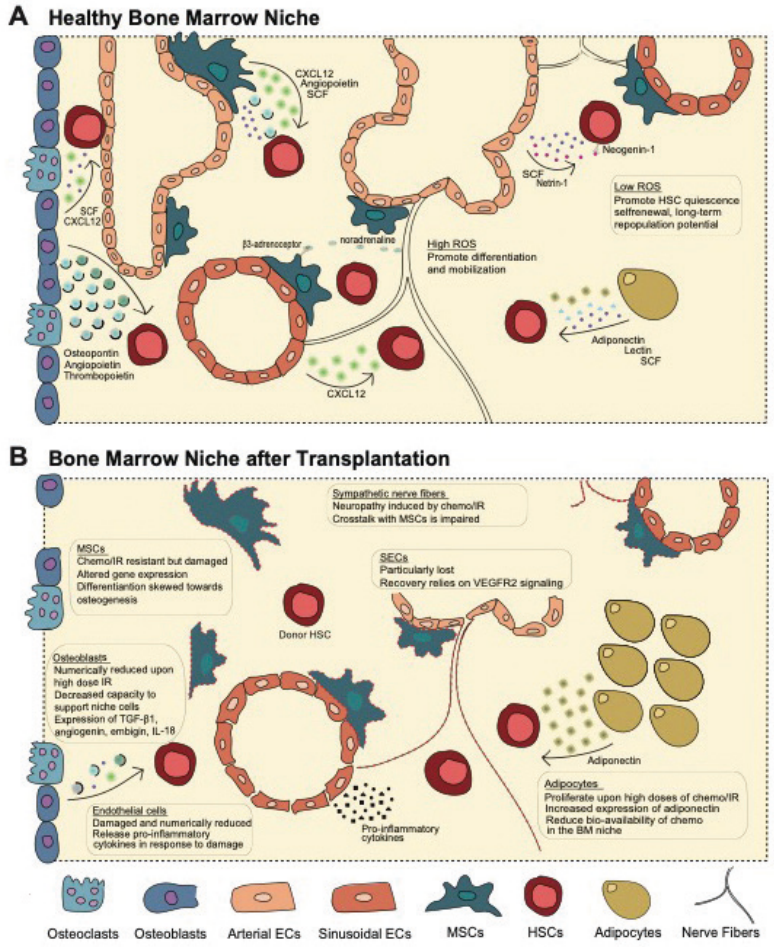


Figure 1: Composition and function of the bone marrow niche after HCT.
Schematic overview of the healthy bone marrow niche (A) and the bone marrow niche after hematopoietic cell transplantation (B). SCF: Stem Cell Factor. CXCL12: C-X-C Motif Chemokine Ligand 12. VEGFR2: Vascular Growth Factor Receptor 1. TGF-β1: Transforming Growth Factor beta 1. ROS: Reactive Oxygen Species. MSCs: Mesenchymal Stromal Cells. ECs: Endothelial Cells. HSCs: Hematopoietic Stem Cells. IR: irradiation.

3

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

Jurrian K. de Kanter[1,2,ǂ], **Flavia Peci** [1,2,ǂ], 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], Mirjam Belderbos[1,2,ǂ,*], Ruben van Boxtel[1,2,$,ǂ,*]

[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] Dept. of Pediatric Oncology/Hematology, Erasmus Medical Center, Rotterdam, 3015 GD, The Netherlands

ǂ These authors contributed equally.

Cell Stem Cell, 2023 DOI: 10.1016/j.stem.2021.07.012

Abstract

Genetic instability is a major concern for the 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 patients who received HSCT or a solid organ transplantation earlier in life. Antiviral treatment with nucleoside analogues can cause enhanced mutagenicity in transplant recipients, which may ultimately contribute to therapy-related carcinogenesis.

Keywords
Hematopoietic stem cell transplantation, somatic mutations, mutational signatures, cancer genomics, antiviral treatment, therapy-related neoplasms, cytomegalovirus, ganciclovir

Introduction

The 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 routinely used for therapeutic purposes. HSC transplantation (HSCT) is performed in >40,000 patients 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 patients 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). Due to increased use of HSCT as a treatment strategy, as well as improved transplantation protocols, the number of HSCT survivors and their life-expectancy continue to increase(Bhatia, 2011). Currently, it is estimated that there are >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, are 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 patients 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 to 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 transplantation. For this, we compared the mutation burden in these cells to 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 patients who underwent renal transplantation earlier in life. These patients 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 analogue ganciclovir, which is administered to immune deficient patients as a first-line treatment of viral reactivation. Our study demonstrates that antiviral treatment with nucleoside analogues post-transplantation can be associated with increased mutagenicity, which may ultimately drive the development of therapy-related malignancies.

Results

Cataloguing 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 catalogue all the mutations that were present in the parental HSPCs (Figure 1A)(Jager et al., 2017; Osorio et al., 2018). We included nine pediatric HSCT recipients, who were transplanted with either bone marrow cells of an HLA-identical sibling donor (n=3, SIB1-3), a haploidentical parent donor (n=2, HAP1-2), or with an anonymous umbilical cord blood (UCB) donor (n=4, CB1-4). 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 per patient 2-14 HSPC clones by WGS, at a depth of 15-30x 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. If a control was unavailable, we used the various clones of the same individual for filtering (see Methods and Table S2). The variant allele frequencies (VAF) 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., 2017; Osorio et al., 2018; Rosendahl Huber et al., 2019). In total, we identified 15691 clonal single base substitutions (SBS) and 927 indels in 51 assessed HSPCs (Table S2-3). We reconstructed phylogenetic trees for all patients 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 artefacts during library preparation or sequencing, we generated new libraries and re-sequenced the genomes of five clones of two patients. In total, we could validate 1049 out of 1070 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 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.

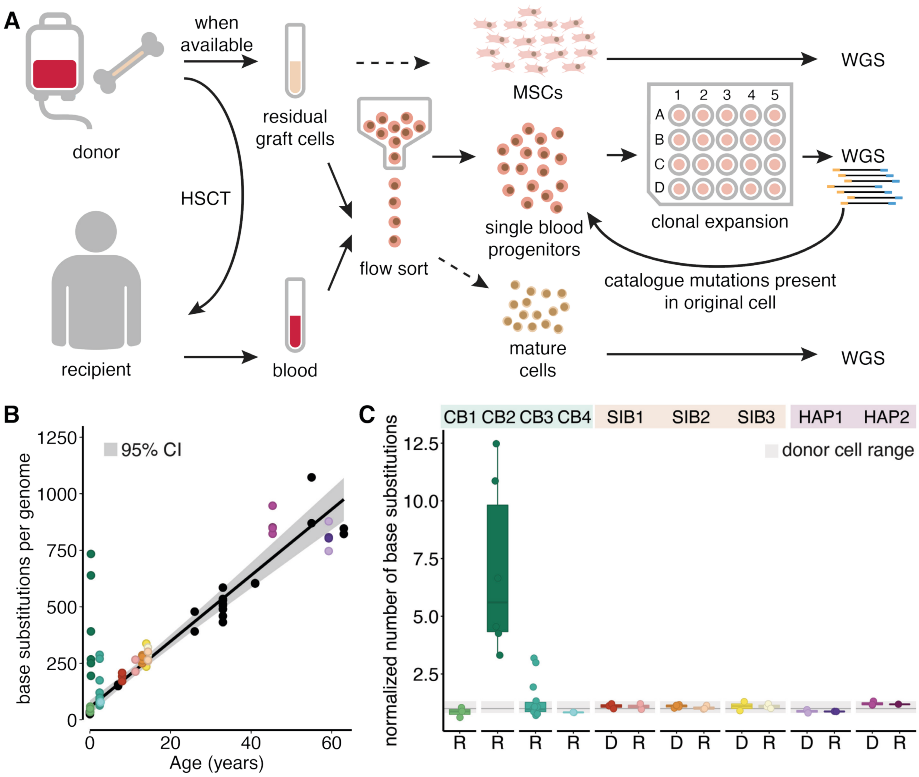


Figure 1: Mutation accumulation associated with HSCT in humans

(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-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 impact of transplantation, we compared the somatic mutation load in HSPCs collected from human HSCT recipients after transplantation to that of their donor's pre-HSCT clones and to this healthy baseline (Figure 1B-C, 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, as 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 ten independent post-HSCT clones with up to twelve-fold more mutations than predicted based on their age (mean observed/expected 5.15, range 1.33-12.5, 95% CI 2.8-7.5; Figure 1B-C), 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 indels was limited and more variable (Figure S1D-F). 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 the post-HSCT clones with a significantly higher base substitution load (Figure S1D-F). Collectively, these data show that, while 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 catalogues 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 the 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 (Figure 2A-C)

(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), of which the underlying mechanism is still unknown. In contrast, in the HSPC clones with an increased number of mutations as compared to the normal baseline, C>A transversions were the most abundant mutation type, accounting for 40-87% of the total number of base substitutions (Figure 2A-D). The number of C>A transversions in these cells was significantly increased as compared to the 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 (Figure 2D, S2), suggesting a single causative process. Indeed, mutational signature analysis revealed that the increase in mutation load in these recipient HSPCs could be exclusively attributed to a previously unidentified single base substitution (SBS) signature, which we called "SBSA" (Figure 2C-D, Table S4). SBSA is characterized by C>A transversions in an NpC>ApA trinucleotide context (86% of all mutations in SBSA), of which >90% are CpC>ApA changes (Figure 2D). SBSA mutations occurred in two out of the nine (22%) assessed patients in this study (CB2, CB3). Of these, 6 out of 6 CB2 clones (100%) and 6 out of 14 CB3 clones (43%) harbored SBSA mutations. To establish if 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 patient CB3. Subsequently, we assessed for each mutation present in the 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). Notably, 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 COSMIC database (v3.0) as well as to *in vitro* established signatures of environmental agents(Kucab et al., 2019; Tate et al., 2019). A cosine similarity of ≥0.95 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; Figure 2F, 4A and S4C).

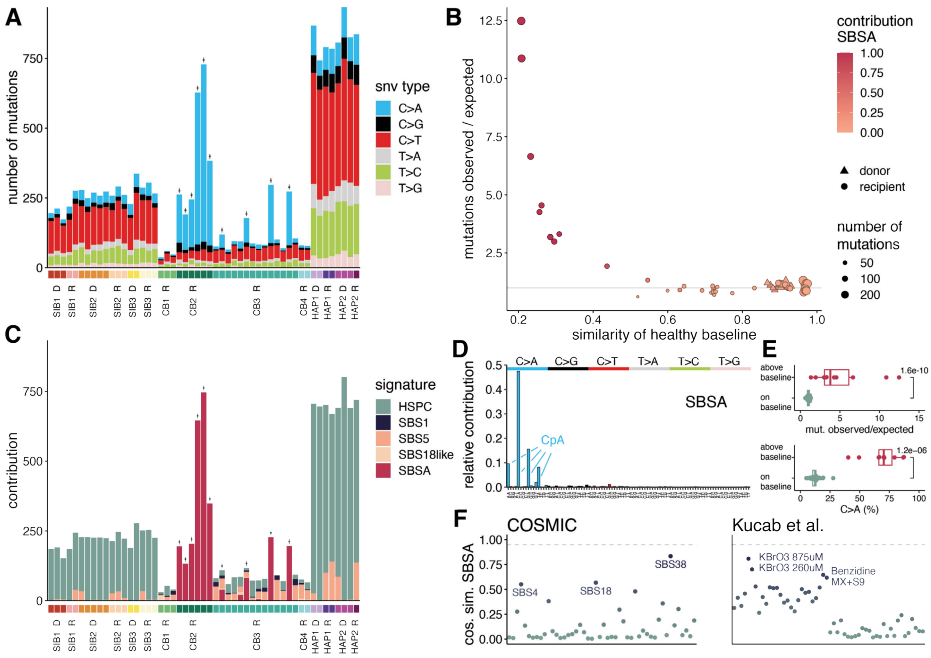


Figure 2: Transplantation-associated mutagenesis can be attributed to a unique mutational signature, SBSA

(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).

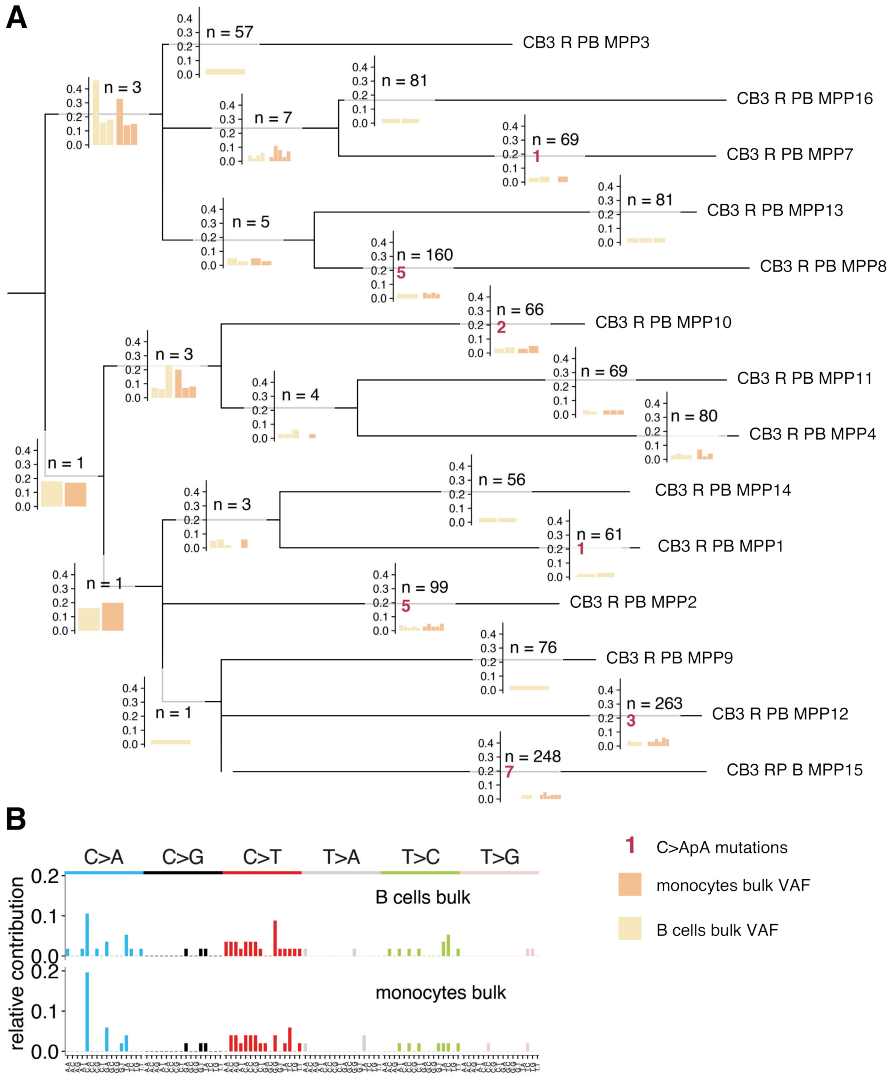


Figure 3: Detection of HSPC mutations in bulk mature populations
(A) The phylogenetic tree of the HSPCs of individual CB3. At each branch, a bar graph is plotted. The number above each bar graph indicates the total number of mutations in that branch. Each bar represents the VAF of a mutation in that 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.

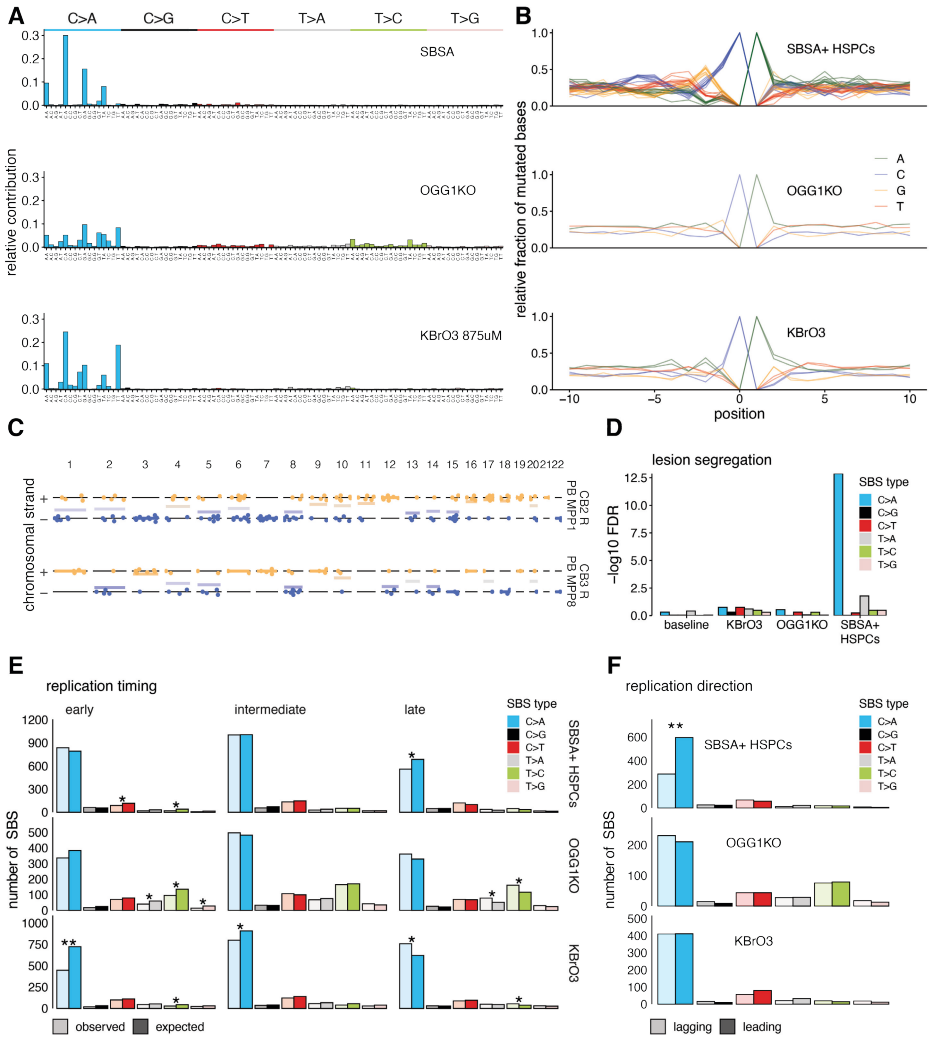


Figure 4: SBSA is characterized by lesion segregation and a strong replication direction bias

(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.

Molecular characterization of SBSA

SBS38, SBS18 and the $KBrO_3$ signature have been attributed to oxidative stress-induced mutagenesis, which is thought to be 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, as 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 to 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) (Figure 4A, figure S4). C>ApA mutations in the 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 (Figure 4B, 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 the post-HSCT clones with high mutation load and contribution of SBSA, the C>A transversions demonstrated a highly significant Watson-versus-Crick-strand lesion segregation (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, 0.48 respectively, Figure 4C-D, S4E). It was previously shown 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., 2020). As a result, one daughter cell and its progeny only carry mutations on either the Watson or the Crick strand, while 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 towards 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 (Figure S4B,F)(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 towards the leading strand (fdr < 10e-23, Figure 4F, S4D), which indicates that the mutagenic process underlying SBSA is directly coupled to DNA replication(Haradhvala et al., 2016; Tomkova et al., 2018). Altogether, these data suggest that, unlike oxidative-stress induced mutations, SBSA mutations in post-HSCT clones are caused by erroneous DNA replication upon a short-term exposure of a mutagenic source.

SBSA is caused by the antiviral nucleoside analogue ganciclovir

To identify the mutagenic source of SBSA, we analyzed the clinical data of our transplant recipients (Table S1). Both HSCT-recipients that 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 dGTP incorporation into DNA(Seley-Radtke and Yates, 2018). FC is a pyrophosphate analogue, which is thought to directly inhibit viral polymerase activity(Crumpacker, 1992). As these compounds affect DNA replication, they are likely candidates for causing SBSA mutations. To test this, we exposed human $CD34^+$ umbilical cord blood HSPCs to GCV and/or FC *in vitro* (Figure 5A). While GCV caused dose-dependent cell death at micromolar concentrations, which are also observed in human plasma (IC50 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 hours with 5μM GCV and/or, similar to previous publications, a 40 times higher concentration of FC (200μM)(Maggs and Clarke, 2004). Both GCV and the combination treatment caused substantial DNA damage, visualized by ɣ-H2AX staining, while FC exposure alone did not cause considerable cell death (Figure 5C,D, 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 single base substitutions as compared to HSPCs exposed to FC alone or untreated clones, with a bias towards C>A transversions (Figure 5E). The number of indels

was similar between GCV, FC and control-treated organoids, no copy number variations or structural rearrangements were found (Figure S5A,B). Importantly, the 96-trinucleotide profile induced by *in vitro* exposure to GCV was essentially identical to SBSA found in human patients (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 towards 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 (Figure S5D-H). Altogether, these data clearly demonstrate that GCV is the cause of the SBSA mutations.

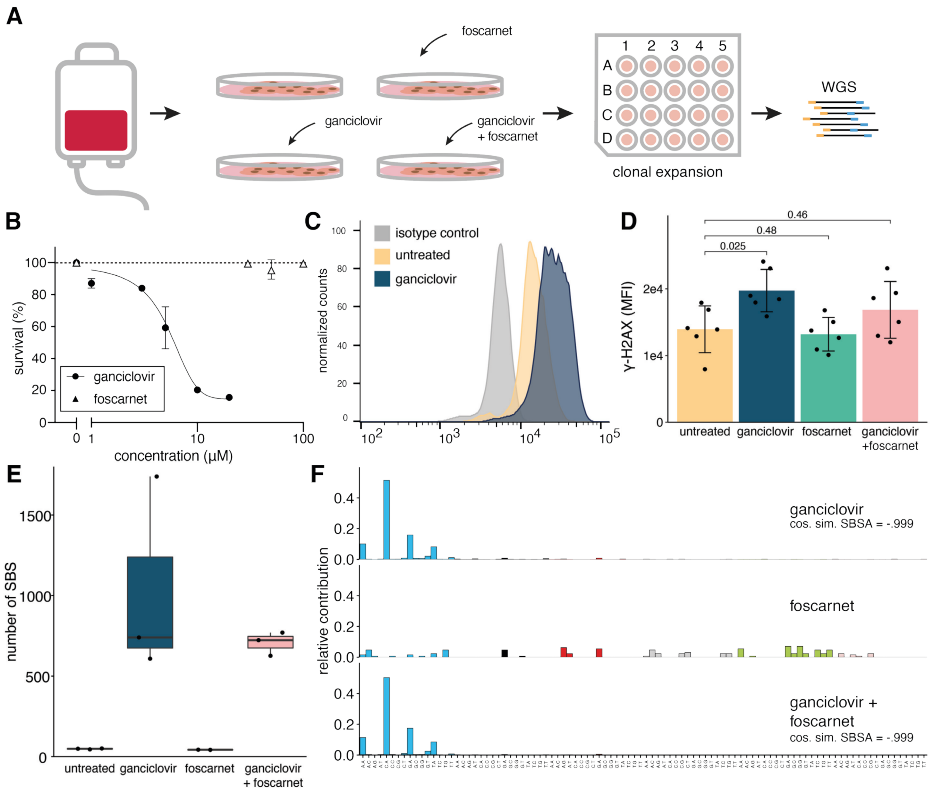


Figure 5: Ganciclovir induces SBSA mutations *in vitro*

(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.

Table 1: Clinical information for SBSA-positive cancers

Sample	Primary diagnosis	Trans-plantation	(second) cancer	Viral reactivations	Antiviral therapy	(Second) cancer driver mutations C>ApA	Ref.
11396 – Dx2 AML	ALL	HSCT	AML	CMV	Ganciclovir, foscarnet,		N/A
633734 – relapse	AML	HSCT	AML-relapse	CMV	Ganciclovir	*NRAS* p.Q61K	(Christopher et al., 2018)
103342 – relapse	AML	HSCT	AML-relapse	CMV	Ganciclovir, valganci-clovir		(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	Donor cell leukemia	Unknown	Unknown	*SETBP1* p.T873K	(Gondek et al., 2016)
CPCT-02090030T	Renal insuffi-ciency	Kidney Tx	Vulvar carcinoma metastasis	Unknown	Unknown	*HRAS*, p.Q61K	(Priestley et al., 2019)
CPCT-02110076T	Renal insuffi-ciency	Kidney Tx	Breast carcinoma metastasis	CMV	Valgan-ciclovir		(Priestley et al., 2019)
CPCT-02340067T	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.

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 single base substitution originates from SBSA, or not (Figure S6A, B, G). We trained the RF on the pre- and post-HSCT HSPCs and on the healthy baseline HSPCs depicted in Figure 1. Importantly, the RF classifier assigned the highest importance to the

nucleotides which 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 1000 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 <1% SBSA-positive mutations in both datasets, confirming the specificity of this classifier.

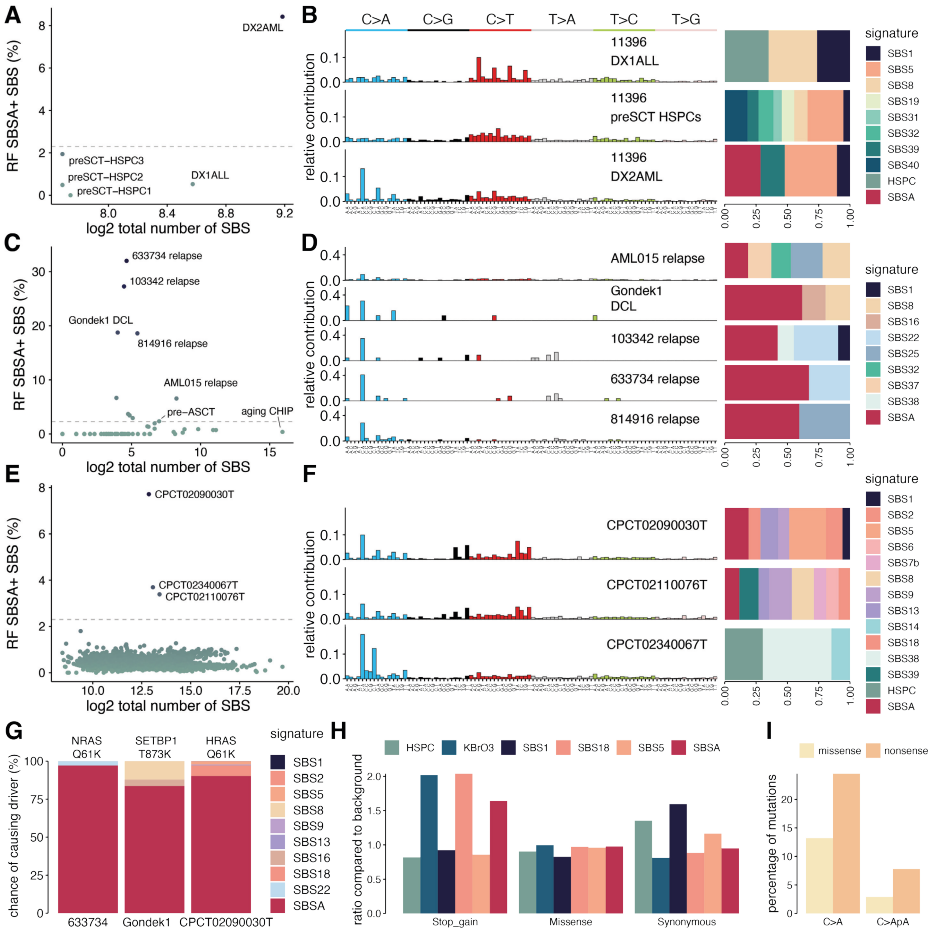


Figure 6: SBSA is present in transplant-related cancers and can cause cancer driver mutations

(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.

Next, we applied this RF classifier to sequencing datasets of human metastatic cancers (n=3668)(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 individual patients as SBSA-positive (Figure 6, Table 1). The first was a therapy-related AML (tAML, PMC11396), in which SBSA had an estimated contribution of 28% (Figure 6A-B). This patient had received an allogeneic HSCT for relapsed acute lymphoblastic leukemia (ALL) with successful engraftment yet developed a tAML of patient-origin three years later (Table 1). Using the RF classifier on WGS data of this patient's tAML, the primary ALL, as well as three normal HSPCs collected three 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 nucleotide context (Figure S6C) and replication strand bias (Figure S6H). The C>A mutations did, however, not display a Watson-versus-Crick bias (Figure S6K). Notably, 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 patient was treated with FC and GCV for a CMV reactivation after HSCT.

The second SBSA-positive tumor was a donor-cell leukemia (DCL), reported in a study by Gondek et al. on clonal hematopoiesis after HSCT and its progression towards malignancy(Gondek et al., 2016)(Figure 6C-D). This patient (called from hereon Gondek1) was transplanted for AML and developed a DCL 3.5 years post-HSCT. The mutation profile of this DCL scored high in the RF (Figure 6C) and had a clear SBSA signature (Figure 6D), while the graft material of this patient, collected before HSCT, did not have these mutations.

Moreover, 4 out of 44 assessed AML relapses were SBSA positive, and all patients had been transplanted (Figure 5C,D)(Christopher et al., 2018). Again, we confirmed this with mutational signature analysis, replication direction bias

and the extended context (Figure S6E,J). Also, in this case, the C>A mutations did not have a Watson-versus-Crick asymmetry (Figure S6K). From 3 out of 4 patients, the medical history could be obtained (Table 1). All three patients developed an early CMV reactivation post-SCT and received GCV as antiviral treatment, consistent with an approximate prevalence of SBSA in 14% (4 out 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 patients had received an HSCT, two out of three patients had received a kidney transplantation earlier in life. For one of these patients, we could retrieve treatment history, which revealed that the patient received GCV to treat a viral reactivation after the transplantation. Further analyses confirmed the SBSA +/-10 nucleotide context and replication strand bias in the metastases of these two transplanted patients, but showed no Watson-versus-Crick asymmetry (Figure 6F, S6D,I,K,L). In contrast, the tumor of the non-transplanted melanoma patient did not show this context nor bias (Figure S6D,I) 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 patients. We estimated the probability of SBSA having caused these mutations using a previously published method(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 damaging potential of SBSA, we calculated the enrichment of stop-gain, missense and synonymous mutations that 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, KBrO3 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 to 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 due to SBSA(Jaiswal et al., 2014). Of these hematologic cancer-drivers, 7.8% of stop-gain mutations were caused by C>ApA mutations, while only 2.8% of non-synonymous mutations occurred in the SBS-context, confirming our previous results (Figure 6H). Taken

together, these results identify the presence of the GCV-induced mutational signature in several types of cancer of human transplantation recipients, and demonstrate its potential to cause cancer driver mutations, in particular stop-gain mutations.

In summary, in this study we provide insight into the impact of HSCT on the acquisition and causative processes of somatic mutations in the transplanted stem cells, and into their impact on malignant transformation. During normal human ageing, HSCs are estimated to acquire 14-15 SNVs per year(Hasaart et al., 2020; Lee-Six et al., 2018). As HSCs divide approximately every 40 weeks(Caitlin et al., 2011), this would mean that if all mutations occur due to stochastic replication errors, each HSC acquires 11 mutations per division. If 1000-5000 transplanted HSCs would repopulate the new blood system and regenerate the estimated average pool of 200.000 HSCs, 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 on 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 from 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 post-transplant 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 patients who had received a kidney transplant previously, supporting the concept that GCV-associated mutagenesis may contribute to the 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 as compared to 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 analogues 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. As GCV is a guanine analogue, 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 pose a possible explanation as to 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. As 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 other signatures than SBSA. This highlights the usefulness of studying pediatric patients, 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, this would occur more efficiently during lagging DNA strand synthesis(Tomkova et al., 2018). However, our data is not definitive proof for this mechanism underlying GCV-induced mutagenesis and the repair of GCV-induced lesions.

GCV is used for the prevention and first-line treatment of CMV disease in transplantation recipients, as well as in patients with congenital CMV infection and CMV reactivation in patients with severe immune deficiency or with HIV/AIDS(Griffiths and Lumley, 2014). Therefore, its mutational consequences are likely to have a more widespread healthcare impact than only in transplantation recipients. The mutagenic effect of GCV and its long-term clinical consequences

should be assessed in large patient 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 analogue 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 that 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 for careful surveillance of HSCT recipients to detect and prevent long-term morbidity.

Limitations of the study

First, although the use of *in vitro* clonal expansion allows to catalogue genome-wide mutations in single HSPCs, it may preferentially select for 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 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 patient cohorts. Similarly, induced

mutagenesis in other tissues needs to be investigated to fully characterize the contribution of this antiviral nucleoside analogue to carcinogenesis. As 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 patient cohort may induce mutations in a subgroup of HSCT recipients.

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. 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.K., F.M., M.J.R., R.O. and R.v.B.; Formal Analysis, J.K.K., M.E.B., M.J.R., R.O. and R.v.B.; Investigation, A.M.B, A.R.H, E.B., M.E.B., F.P., A.v.L; Writing – Original Draft, M.E.B., J.K.K. and R.v.B.; Supervision, M.B. and R.v.B; Funding acquisition, M.E.B. and R.v.B.

Declaration of interests

The authors declare no competing interests.

References

1. Abascal, F., Harvey, L.M.., Mitchell, E., Lawson, A.R.., 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.
2. Aitken, S.S.J., Anderson, C.J., Connor, F., Pich, O., Sundaram, V., Feig, C., Rayner, T.F., Lukk, M., Aitken, S.S.J., Luft, J., et al. (2020). Pervasive lesion segregation shapes cancer genome evolution. Nature ***583***, 265–270.
3. 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.
4. 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.
5. 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. (2013). Signatures of mutational processes in human cancer. Nature ***500***, 415–421.
6. 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 (80-.). ***354***, 618–622.
7. Andrews, P.W., Ben-David, U., Benvenisty, N., Coffey, P., Eggan, K., Knowles, B.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.
8. Avior, Y., Eggan, K., and Benvenisty, N. (2019). Cancer-related mutations identified in primed and naive pluripotent stem cells. Cell Stem Cell ***25***, 456–461.
9. 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.
10. 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.
11. 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.
12. 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.
13. Bhatia, S. (2011). Long-term health impacts of hematopoietic stem cell transplantation inform recommendations for follow-up. Expert Rev. Hematol. ***4***, 437–454.
14. 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.
15. 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. (2020). Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature ***586***, 763–768.
16. 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.

17. Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. ***10***, 33.
18. Boettcher, S., Wilk, C.M., Singer, J., Beier, F., Burcklen, E., Beisel, C., Ferreira, M.S.V.V., 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.
19. 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.
20. 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***, 1–9.
21. Burns, S.S., and Kapur, R. (2020). Stem Cell Reports Review Clonal Hematopoiesis of Indeterminate Potential as a Novel Risk Factor for Donor-Derived Leukemia. Stem Cell Reports ***15***, 279–291.
22. Caitlin, S.N., Busque, L., Gale, R.E., Guttorp, P., Abkowitz, J.L., 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.
23. 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.
24. 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. U. S. A. ***111***, 17462–17467.
25. 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.
26. Clark, C., Savani, M., Mohty, M., and Savani, B. (2016). What do we need to know about allogeneic hematopoietic stem cell transplant survivors? Bone Marrow Transplantation2 ***51***, 1025–1031.
27. Collins, F.S., and Gottlieb, S. (2018). The next phase of human gene-therapy oversight. N. Engl. J. Med. ***379***, 1393–1395.
28. Crumpacker, C.S. (1992). Mechanism of action of foscarnet against viral polymerases. Am. J. Med. ***92***, S3–S7.
29. 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***, D794–D801.
30. 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 nextgeneration DNA sequencing data. Nat. Genet. ***43***, 491–498.
31. Dunbar, C.E., High, K.A., Joung, J.K., Kohn, D.B., Ozawa, K., and Sadelain, M. (2018). Gene therapy comes of age. Science (80-.). ***359***, 1–10.
32. 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***, D766–D773.
33. 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.

34. Griffiths, P., and Lumley, S. (2014). Cytomegalovirus. Curr. Opin. Infect. Dis. *27*, 554–559.
35. 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.
36. Haradhvala, N.J., Polak, P., Koren, A., Lawrence, M.S., and Getz Correspondence, G. (2016). Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell *164*, 538–549.
37. Hasaart, K., Manders, F., van der Hoorn, M.-L., Verheul, M., Poplonski, T., Kuijk, E., Chuva de Sousa Lopes, S., and van Boxtel, R. (2020). Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis. Sci. Rep. *10*, 12991.
38. 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.
39. 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.
40. Jager, M., Blokzijl, F., Sasselli, V., Boymans, S., Janssen, R., Besselink, N., Clevers, H., van Boxtel, R., and Cuppen, E. (2017). Measuring mutation accumulation in single human adult stem cells by whole-genome sequencing of organoid cultures. Nat. Protoc. *13*, 59–78.
41. 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.
42. Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, 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.
43. 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*, 1–12.
44. 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.
45. 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.
46. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.
47. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
48. Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. R News *2*, 18–22.
49. 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. U. S. A. *96*, 1492–1497.

50. 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.
51. 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. U. S. A. *116*, 1447–1456.
52. Lüdecke, D. (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. J. Open Source Softw. *3*, 772.
53. 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.
54. Majhail, N.S., Tao, L., Bredeson, C., Davies, S.M., 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 Transplantation2 *19*, 1498–1501.
55. 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.
56. 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*, 1–12.
57. 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*, 1–11.
58. 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*, 1–9.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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. (2020). Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature *580*, 269–273.

65. Priestley, P., Baber, J., Lolkema, M., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidairi, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumors. Nature ***575***, 210–216.
66. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics ***26***, 841–842.
67. R Core Team (2020). R: A Language and Environment for Statistical Computing.
68. 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***, 1–11.
69. Rosendahl Huber, A., Manders, F., Oka, R., and van Boxtel, R. (2019). Characterizing Mutational Load and Clonal Composition of Human Blood. JoVE.
70. Scala, S., Basso-Ricci, L., Dionisio, F., Pellin, D., Giannelli, S., Salerio, F.A., Leonardelli, L., Cicalese, M.P., Ferrua, F., Aiuti, A., et al. (2018). Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. Nat. Med. ***24***, 1683–1690.
71. 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.
72. 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.
73. Stunnenberg, H.G., Abrignani, S., Adams, D., de Almeida, M., Altucci, L., Amin, V., Amit, I., Antonarakis, S.E., Aparicio, S., Arima, T., et al. (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell ***167***, 1145–1149.
74. 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.
75. 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***, D941–D947.
76. 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***, 1–14.
77. Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Br. Bioinforma. ***14***, 178–192.
78. 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.
79. Weiner, J. (2017). riverplot: Sankey or Ribbon Plots.
80. Weissman, I.L. (2000). Stem cells: Units of development, units of regeneration, and units in evolution. Cell ***100***, 157–168.
81. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).
82. 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.

83. Yamanaka, S. (2020). Pluripotent stem cell-based therapy - Promise and challenges. Cell Stem Cell *27*, 523–531.
84. LiftOver - USCS Genome Browser.

STAR⋆Methods

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, accession number 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 Table 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 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[21]. The following combinations of cell surface markers were used to define cell populations[47]: 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-γH2A.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-γH2A.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 x 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(Thorvaldsdottir 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)[20]. 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., 2017). 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.

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(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(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(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[45]. 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 clonesand three clones with both 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 Nat Commun. 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([CSL STYLE ERROR: reference with no printed form.]).

Mutation datasets

The data of a knock-out of ***OGG1*** in the human neuroblastoma cell line CHP134 was courteously provided by Jan Molenaar (van den Boogaard et al., under submission). 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 & Kapur(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.(Lee-Six et al., 2018). The AML relapse data were obtained from Christopher et al.(Christopher et al., 2018) and Stratmann et al.(Stratmann et al., 2021). Data on the post-HSCT neoplasms were obtained from Berger et al.(Berger et al., 2018) and Gondek et al(Gondek et al., 2016). The authors of Stratmann et al. 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 artefacts signatures. Each sample for which these artefacts contributed more than 20% were excluded from further analyses. Mutations were transferred to hg19 using UCSC's liftOver([CSL STYLE ERROR: reference with no printed form.]). The aging CHIP dataset was obtained from Bick et al(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[21]. 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 using the ggplot2 package in R(Wickham, 2016).

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 (www.trialregister.nl).

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.

SUPPLEMENTARY INFORMATIONS

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

Jurrian K. de Kanter, Flavia Peci, Eline Bertrums, Axel Rosendahl Huber, Anaïs van Leeuwen, Markus J. van Roosmalen, Freek Manders, Mark Verheul, Rurika Oka, Arianne M. Brandsma, Marc Bierings, Mirjam Belderbos, and Ruben van Boxtel

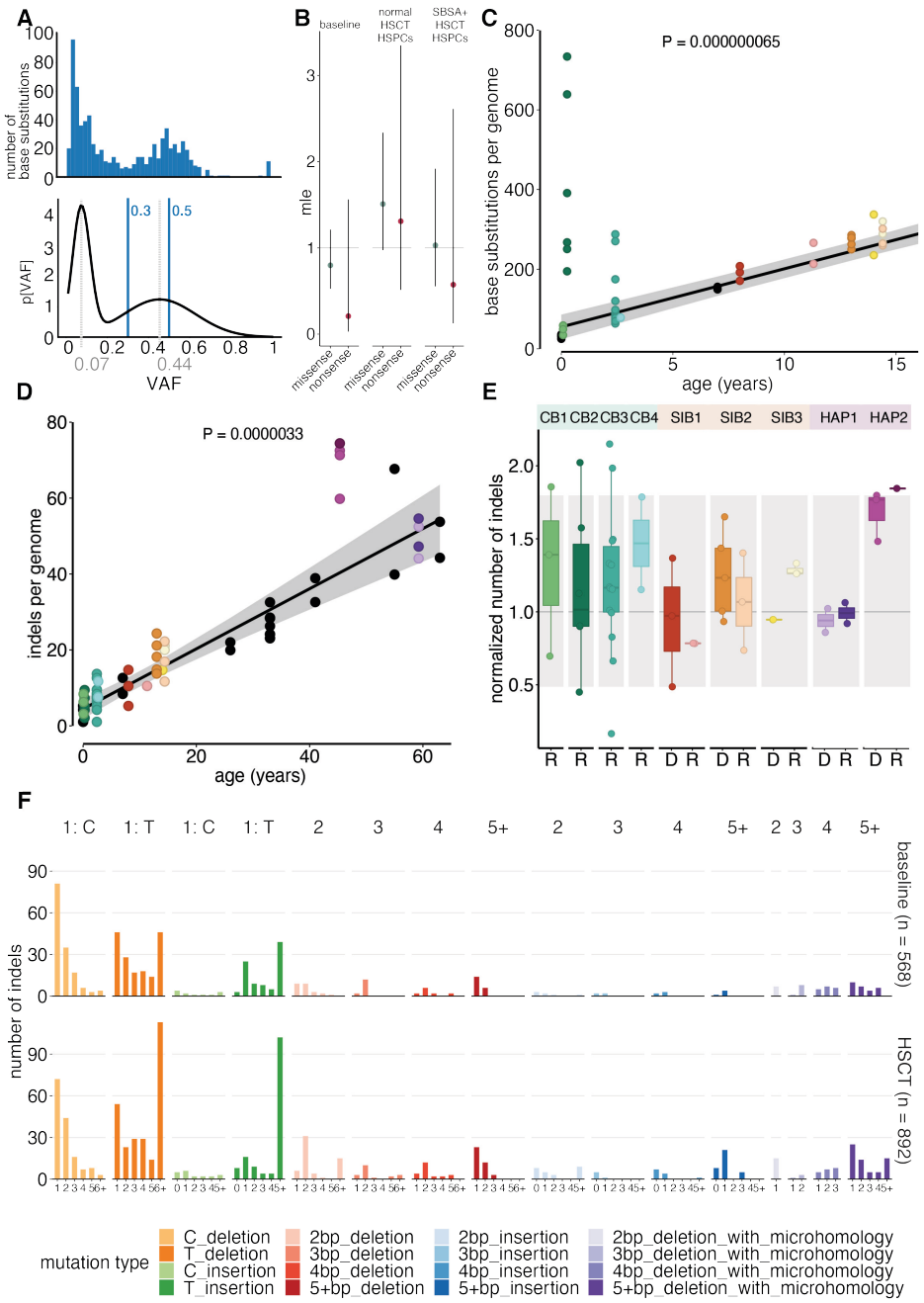


Figure S1. The number of indels in HSCT donor and recipient HSPCs is variable but not consistently altered after transplantation, related to Figure 1.

(A) A representative VAF plot of a HSPC clone. Above, a histogram of the variant allele

frequency (VAF). Below, the probability distribution of these VAFs is shown, with the peaks of the subclonal (0.07) and clonal mutations (0.44) highlighted. (B) dn/ds analysis of nonsynonymous (either missense or nonsense) versus synonymous mutations. "mle" is the maximum likelihood estimate of the ratio between the nonsynonymous and synonymous mutations. (C) Similar to Figure 1B, but zoomed into the HSPC clones of pediatric donors and recipients. The corresponding P value of the linear mixed-effects model of the baseline is depicted above the baseline. (D) Similar to Figure 1B, but the number of indels are shown instead of the number of base substitutions for all samples and the baseline. (E) The number of indels per clone normalized to the indel baseline, similar to Figure 1C. (F) Indel context profiles from the baseline and the HSCT clones.

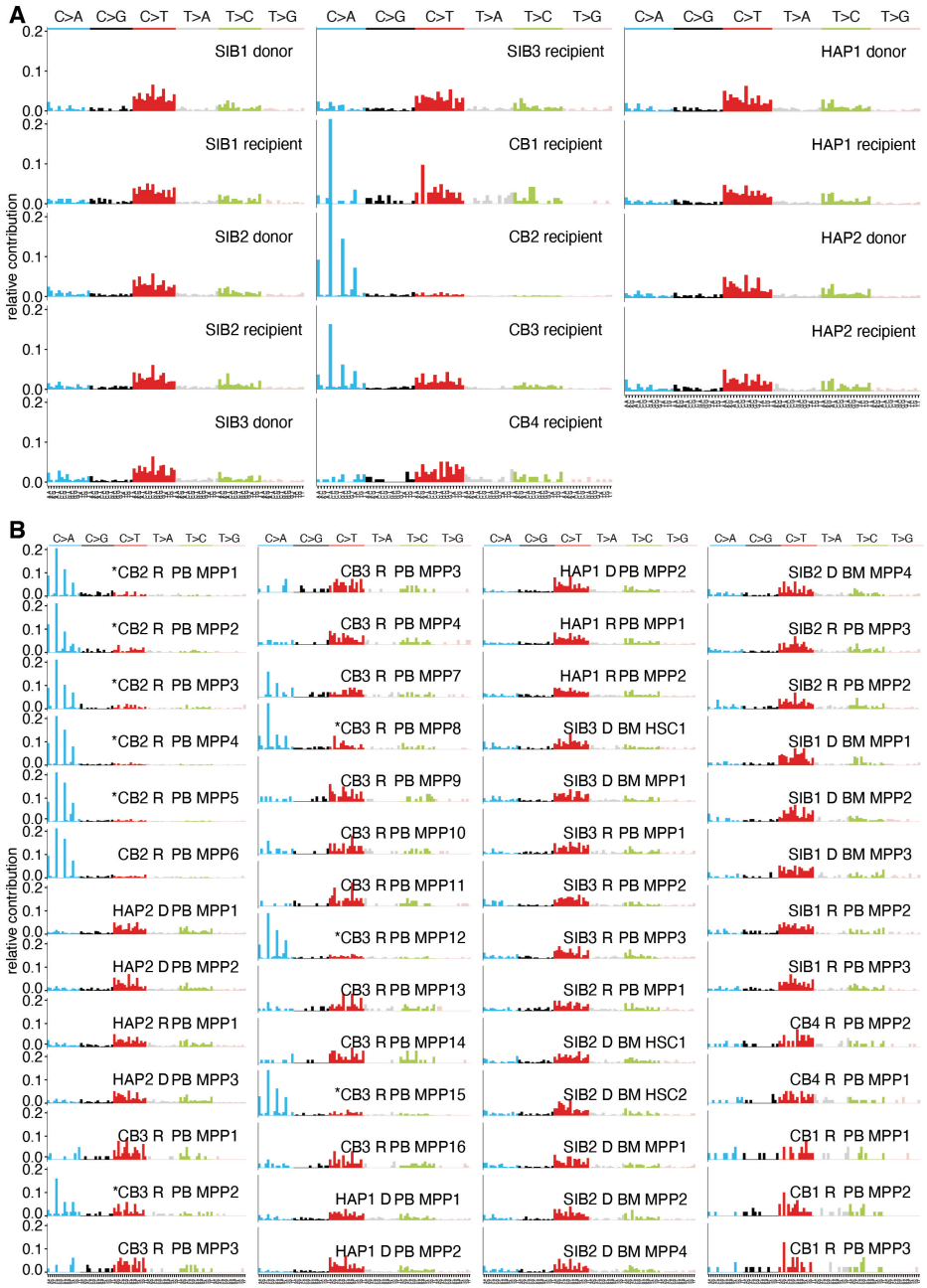


Figure S2. Verification of WGS results by phylogenetic analyses and re-sequencing, related to Figure 2.

(A) Phylogenetic analysis of the clones per patient. The trees indicate which mutations are

shared between clones, and which mutations are only present in individual mutations. Most mutations are acquired in single clones. (B) Re-sequencing of DNA of five HSPC clones from two patients, CB2 (n=2) and CB3 (n=3). Above each bar, the number of mutations identified in the original sequencing of the clone and the number of these mutations that are found at a VAF of 0.15 or higher in the re-sequenced sample are shown.

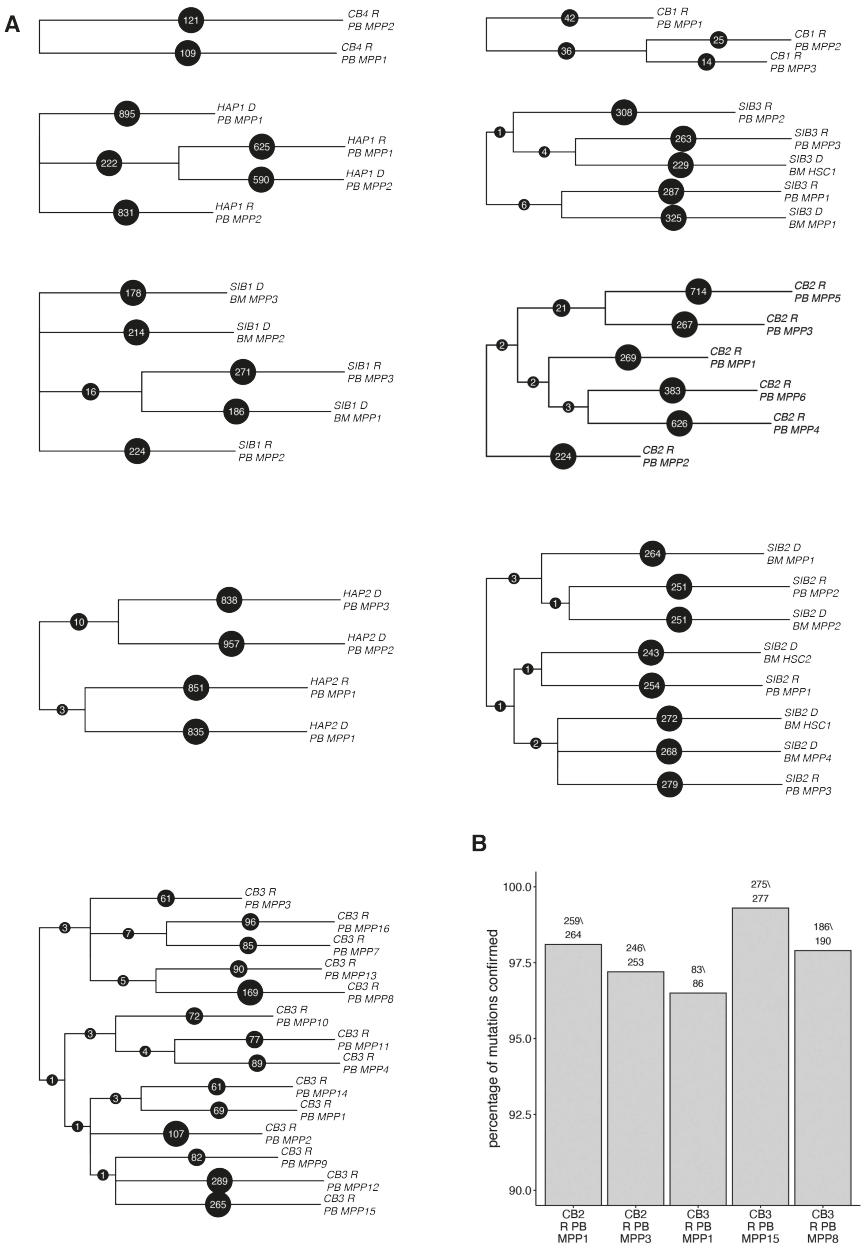


Figure S3. HSCT recipient clones with increased mutational burden have higher contribution of NpC>ApA mutations, related to Figure 3.

(A) SBS 96-trinucleotide profiles of HSPC clones, summed per patient and donor and recipient origin. (B) SBS 96-trinucleotide profiles of all 51 individual HSCT HSPC clones in this study. Names of clones with increased mutation burden are indicated by a "*".

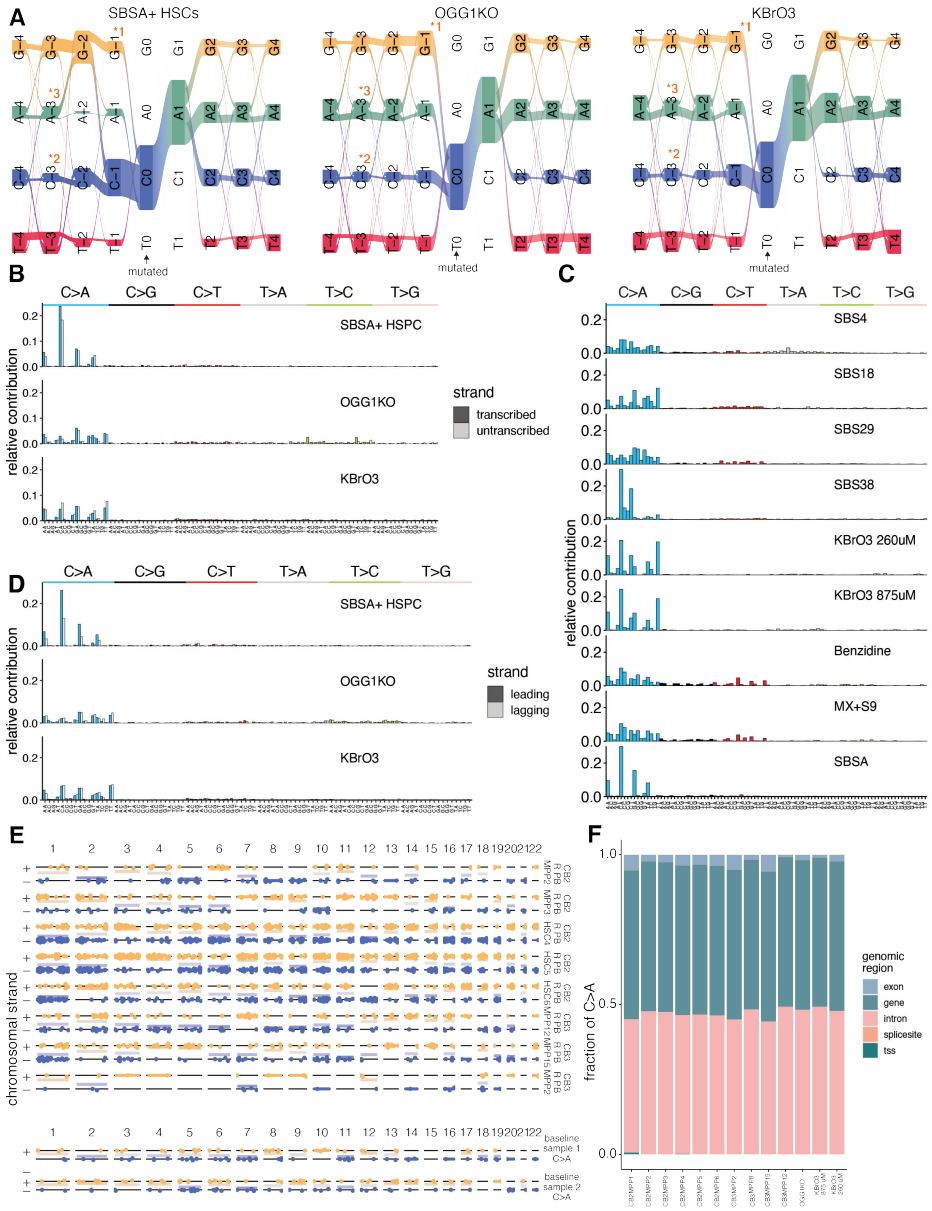


Figure S4. SBSA has a distinct nucleotide context and other characteristics that are distinct from previously reported signatures, related to Figure 4.

(A) Riverplots indicating the order of the -4:+4 nucleotide context of C>ApA mutations of SBSA positive HSPCs, a knock-out of OGG1 and exposure to potassium bromate (KBrO3). The mutated C is present on position 0. SBSA C>A mutations have increased G-2 preceding G-1 (*1), C-2 following C-3 (*2) and decreased A-3 preceding C-2 (*3). (B) The 96-trinucleotide context separated by the transcribed and untranscribed

strand op protein-coding genes. Samples are the same as those depicted in A. (C) The 96-trinucleotide context of COSMIC and environmental agent mutational signatures with the highest correlation to SBSA as shown in Figure 2E. (D) The 96-trinucleotide context separated by the leading and lagging strand of replication for the same samples as those depicted in A. (E) The chromosomal strand of C>A mutations of HSCT recipient HSPC clones with increased mutational burden not shown in Figure 3C, and of two baseline HSPC clones. (F) The distribution of C>A mutations over functional genomic regions of HSCT recipient clones with increased mutation burden, a knock-out of OGG1 and KBrO3 treated cells.

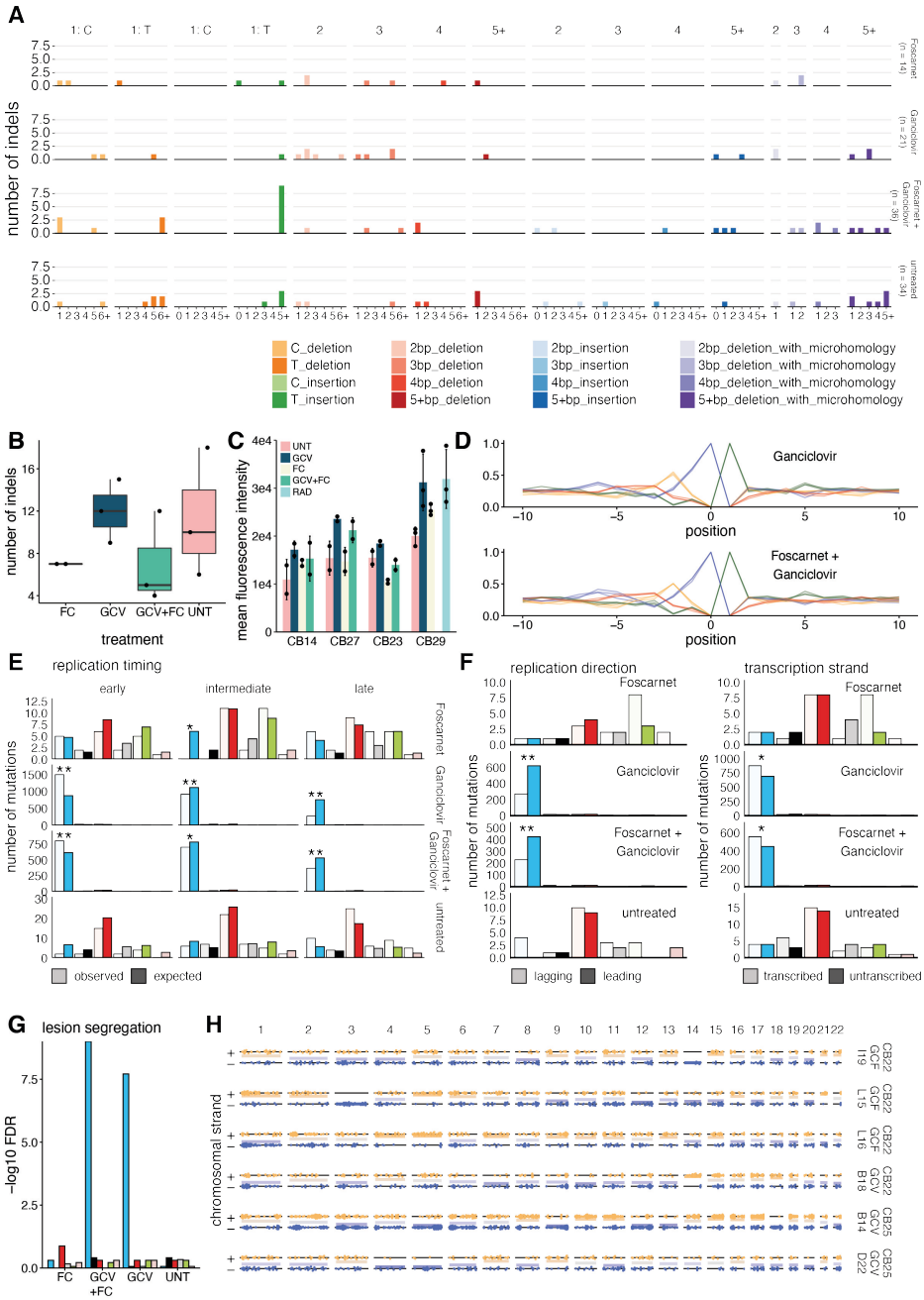


Figure S5. Mutations induced *in vitro* by ganciclovir have similar molecular characteristics as SBSA, related to Figure 5.

Molecular characterization of *in vitro* treatment of umbilical human cord blood cells

with 5 μM Ganciclovir, 200μM Foscarnet, or a combination of both. * = FDR <0.05, ** = FDR < 10-7. (A) The indel context profiles of the clones of each treatment condition. (B) The number of indels per treatment condition. Each dot represents a single clone. (C) The mean fluorescence intensity, similar to figure 4C, but grouped per cord blood sample, and including CB29, for which radiation treatment was available, but not the combined treatment of foscarnet and ganciclovir. (D) The -10:+10 nucleotide context of ganciclovir and a combination of ganciclovir and foscarnet. (E) Enrichment/depletion of the clones from each treatment condition divided in early, intermediate and late replicating regions. Data from clones of one condition were pulled. (F) Replication strand bias and transcription strand bias of the same data as depicted in E. (G) FDR-corrected p-values of Wald-Wolfowitz runs test on summed numbers of mutations and runs in each treatment condition. (H) The chromosomal strand and position of the cytosine of C>A mutations for all clones of all treatment conditions. Abbreviations: FC = foscarnet, GCV = ganciclovir, UNT = untreated, RAD = radiation.

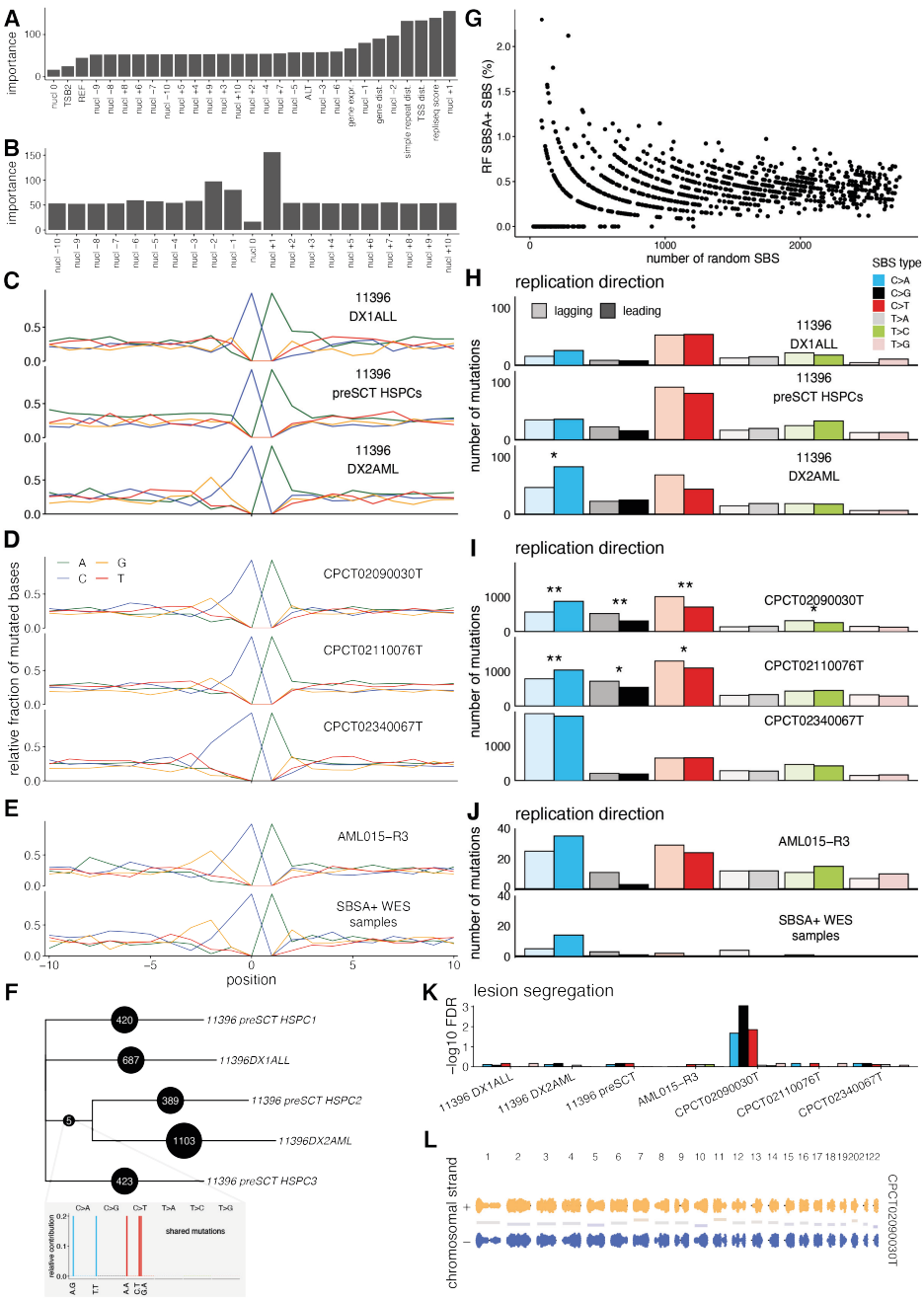


Figure S6. A random forest-based approach identifies tumors with contribution of SBSA to their mutational profile, related to Figure 6.

(A) The importance given by the random forest to the mutation characteristics sorted

from low to high. For each mutation the +10:-10 nucleotide context was used, the Repliseq score, distance to the closest TSS, gene body, and simple repeat, reference (REF) and alternative (ALT) allele and transcriptional strand bias (TSB2). (B) Similar to A, but only the importance of the nucleotide context is shown, sorted by position. (C) The -10:+10 nucleotide context of the C>ApA mutations of the primary (DX1) ALL, pre-HSCT HSPC clones (pulled) and therapy-related (DX2) AML of patient PMC11396. (D) Similar to C, but for the three samples classified SBSA positive by the random forest of the Dutch solid tumor metastases dataset. (E) Similar to C, but for AML015-R3, an AML relapse after allogeneic-HSCT, and the merged data of four SBSA+ classified WES samples. (F) The developmental lineage tree of the samples of patient PMC11396, based on shared mutations. The nucleotide context of mutations shared between the secondary (DX2) AML and HSPC3 is shown. (G) The number of SBS and percentage of random forest SBSA-positively classified SBS of 1000 sets of randomly sampled mutations. The highest percentage (2.3%) was used as a cut-off for expected false-positive rates for input samples. (H), (I), (J) The replication strand bias of the samples in C,D and E respectively. * = FDR <0.05, ** = FDR < 10-7. (K) FDR-corrected p-values for Wald-Wolfowitz runs test for the same samples as C, D and E, similar to Figure 3D. (L) The chromosomal strand of the cytosines of C>A mutations for CPCT02090030T.

Supplementary Table S3. The somatic base substitutions and indels called in HSCT recipient and donor HSPC clones.*

Supplementary Table S4. The 96-trinucleotide mutation profile of SBSA.*

*Extended tables are available at DOI:https://doi.org/10.1016/j.stem.2021.07.012

4

Assessing the mutagenic impact of genotoxins in human umbilical cord blood-derived stem and progenitor cells

Anaïs J. C. N. van Leeuwen[1,3,†], Axel Rosendahl Huber[1,3,†], **Flavia Peci**[1], Jurrian K. de Kanter[1], Eline J.M. Bertrums[1,2], Ruben van Boxtel[1,4,*]

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

[2] *Department of Pediatric Oncology, Erasmus Medical Center, 3015 GD Rotterdam, the Netherlands*

[†] These authors contributed equally

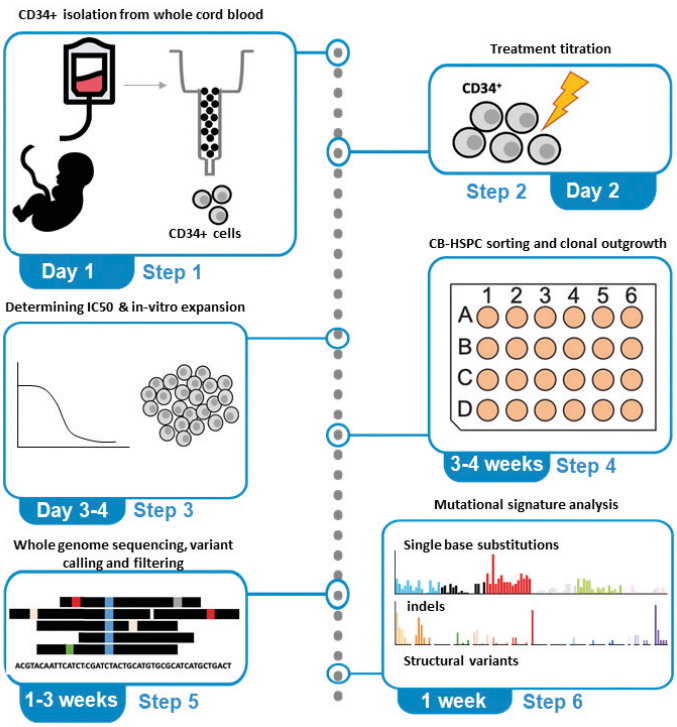
STAR Protocols, DOI: 10.1016/j.xpro.2022.101361

Summary

Many different mutational signatures have been identified in cancer genomes (Alexandrov et al., 2020), providing information about the causes of cancer and treatment vulnerabilities (Helleday et al., 2014). However, for many signatures the underlying cause remains uknown. This protocol describes an approach to determine mutational patterns induced by genotoxins, based on whole-genome sequencing of cord-blood derived hematopietic stem and progenitor cells (CB-HSPCs). The low mutation background of these CB-HSPCs allows for a sensitive identification of additional mutagenic processes.

For complete details on the use and execution of this protocol, please refer to Kanter et al., Cell Stem Cell, 2021 (de Kanter et al., 2021).

Graphical abstract



Before you begin

This protocol describes an efficient method to investigate the mutational consequences of genotoxic exposure in the genomes of human umbilical cord blood-derived hematopoietic stem and progenitor cells (CB-HSPCs). These CB-HSPCs are obtained by enriching for $CD34^+$ cells from either freshly harvested umbilical cord blood or cryopreserved cord blood mononuclear blood cells (CMBCs). These cells are ideally suited to assess mutagenicity of environmental exposure, as CB-HSPCs have a very low background mutation load, containing on average 30.8 unique single base substitutions (SBS) and 4.5 short insertions and deletions (indels) genome-wide. In addition, the human hematopoietic system is very sensitive to genotoxic exposure and, in fact, dose-limiting to the toxic effects of chemotherapy (Crawford et al., 2004). Therefore, this cell system is highly relevant for human toxicity screens. The duration of the protocol is six to nine weeks, including four days for culture initiation and exposure, three to four weeks for clonal expansion, and two to four weeks for WGS and mutation analysis. For the CB-HSPC culture method and the WGS data analysis, a previously published protocol to study mutation load during life (Rosendahl Huber et al., 2019) was adopted, and optimized, by supplementing the CB-HSPC culture media with two components and culturing at hypoxic conditions to improve clonal outgrowth. In addition, it is highly reccommended to repeat the experiment with material of at least two individual cord blood donors, to ensure reproducibility in two individual backgrounds. The protocol is divided into two preparatory steps, followed by six main steps:

1) Preparing CMBCs
2) Isolating $CD34^+$ cells from CBMCs
3) Exposure of $CD34^+$ cells to genotoxic compound(s)
4) HSPC sort and clonal expansion
5) Harvesting HSPC clones, DNA extraction and WGS
6) Variant calling and variant filtration
7) Analysis of *in vitro* induced somatic mutations

Preparation one: Bioinformatic analyses

To perform bioinformatic analyses, we make use of a raw sequencing data processing pipeline based on the Genome Analysis Toolkit (GATK) best practices for data pre-processing and genotyping using the GATK HaplotypeCaller (NF-IAP). To filter variants and obtain high-quality somatic mutations, we employ a python script (SMURF). To visualize mutation data and compare mutational profiles, we make use of the R bioconductor package 'MutationalPatterns'. The steps below describe how to install the bioinformatic tools required for the analysis.

1. Install the NF-IAP pipeline on a computing cluster running SLURM (recommended) or SGE as the workload manager. Detailed instructions for installation can be found at https://github.com/ToolsVanBox/NF-IAP.
2. To obtain high-quality somatic mutations, we apply stringent filters using a Somatic Mutation Rechecker and Filtering (SMuRF) script, available on GitHub. See guidelines for usage on: https://github.com/ToolsVanBox/SMuRF.
   a. Download the mutation blacklist on https://github.com/ToolsVanBox/Genotoxin_assay and set the right path in the config.ini file.

3. To detect structural variants, we make use of the standalone GRIDSS-PURPLE-LINX pipeline (Cameron et al., 2019), which can be retrieved from https://github.com/hartwigmedical/gridss-purple-linx. In the case of using a high-performance computing cluster, using singularity is recommended.

```
$ singularity pull docker://gridss/gridss-purple-linx
```

4. To perform mutational signature analysis, first install R version 4.1.0 or higher from https://www.r-project.org/, if not already installed. After the installation of R, install the 'MutationalPatterns' package using the 'biocManager' package as instructed on the bioconductor website: https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html.

```
> if (!requireNamespace("BiocManager", quietly = TRUE))
+ install.packages("BiocManager")

> BiocManager::install("MutationalPatterns")
```

To detect differences in mutation load, a minimal number of exposed and control clones needs to be sequenced. The number of clones required to determine a significant increase in mutation load depends on the absolute number of induced mutations, and the variation in mutation loads. Before obtaining any result, it is impossible to determine the number of clones to sequence. However, data from an initial CB-HSPC sequencing project can be used to estimate the number of clones to sequence. We have provided a script (on https://github.com/ToolsVanBox/Genotoxin_assay), which takes input from previously sequenced exposed and control clones. This script gererates mutation numbers based on the mean and standard deviation from input conditions (control and exposed condition), and tests different numbers of control and exposure clones. Here, we can determine that for 2 Gy irradiation, the sequencing of four clones for both control and 2 Gy exposure conditions is required to determine a statistical significant increase in mutations (Figure 1). See below an example, using a list of SBS mutation numbers of control-exposed named 'ctrl_muts' and SBS numbers from 2 Gy irradiated CB-HSPCs named 'cond_muts',

```
source("est_n_clones.R)
est_n_clones(ctrl_muts, cond_muts)
```

Preparation two: Preparation of HSPC media

1. Thaw a bottle of 500 mL HSPC media during 24 hours at 4°C. Aliquot into aliquots of 25 mL, freeze at -20°C, and do not expose to freeze-thaw cycles.

Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Human BV421-CD34 (clone 561)	BioLegend	Cat#343609, RRID: AB_11147951
Human FITC-Lineage mix (CD3/14/19/20/56)	BioLegend	Cat#348701, RRID: AB_10644012
Human PE-CD38 (clone HIT2)	BioLegend	Cat#303505, RRID: AB_314357
Human APC-CD90 (clone 5E10)	BioLegend	Cat#328113, RRID: AB_893440
Human PerCP/CY5.5-CD45RA (clone HI100)	BioLegend	Cat#304121, RRID: AB_893358
Human Pe/Cy7-CD49f (Clone GoH3)	BioLegend	Cat#313621, RRID:
Human FITC-CD16 (clone 3G8)	BioLegend	Cat#302005, RRID: AB_314205
Human FITC-CD11c (clone 3.9)	BioLegend	Cat#301603, RRID: AB_314173
Bacterial and virus strains		
N/A	N/A	N/A
Biological samples		
Human cord blood	N/A	N/A
Chemicals, peptides, and recombinant proteins		
Human Stem Cell Factor (SCF)	Miltenyi Biotec	Cat#130-096-696
Human Flt3 – Ligand	Miltenyi Biotec	Cat#130-096-480
Interleukin-6 (IL-6)	Stem Cell Technologies	Cat#78050
Interleukin-3 (IL-3)	Stem Cell Technologies	Cat#78040
Thrombopoietin (TPO)	Miltenyi Biotec	Cat#130-095-754
Primocin	InvivoGen	Cat#ant-pm-05
UM729	Stem Cell technologies	Cat#72332
Stemregenin-1 (SR1)	Stem Cell Technologies	Cat#72344
Critical commercial assays		
CD34 Microbead Kit UltraPure, human	Miltenyi Biotec	Cat#130-100-453
LS Columns	Miltenyi Biotec	Cat#130-042-401

Key resources table (Contintued)

REAGENT or RESOURCE	SOURCE	IDENTIFIER
QIAamp DNA Micro Kit	Qiagen	Cat#56304
TruSeq DNA Nano	Illumina	Cat#20015964
Deposited data		
Whole-genome sequencing data of exposed cord blood cells	-	EGA
Experimental models: Cell lines		
Experimental models: Organisms/strains		
N/A	N/A	N/A
Oligonucleotides		
N/A	N/A	N/A
Recombinant DNA		
N/A	N/A	N/A
Software and algorithms		
Nextflow Illumina Analysis Pipeline (NF-IAP)	https://github.com/ToolsVanBox/NF-IAP	Version 1.3.0
GRIDSS-PURPLE-LINX standalone pipeline	https://github.com/hartwigmedical/gridss-purple-linx	Version 1.3.2
Somatic Mutations Rechecker and Filtering (SMuRF)	https://github.com/ToolsVanBox/SMuRF	Version > 2.0.0
R	https://www.r-project.org/	Version > 4.0.0
R package: MutationalPatterns	https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html	Version > 3.0.0
R package: drc: Analysis of Dose-Response Curves	https://cran.r-project.org/web/packages/drc/index.html	Version > 3.0.0
Supplementary code for displaying mutation contexts and mutation blacklists	https://github.com/ToolsVanBox/Genotoxin_assay	
Other		

Key resources table (Contintued)

REAGENT or RESOURCE	SOURCE	IDENTIFIER
50 mL LeucoSep tube with porous insert, pre-filled with Leucosep separation medium	Greiner	Cat#227288
Lymphoprep	Stem Cell Technologies	Cat#07861
StemSpan SFEM I Medium	Stem Cell Technologies	Cat#09650
Iscove's Modified Dulbecco's Medium (IMDM)	Thermo Fisher Scientific	Cat#12440053
CELLSTAR plate, 384w, 130 µL, F-bottom, TC, cover	Greiner	Cat#781182
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich	Cat#D8418, CAS: 67-68-5
Fetal bovine serum (FBS)	ThermoFisher	Cat#10500
Bovine serum albumin (BSA)	Sigma-Aldrich	Cat#A9647-50G, CAS: 9048-46-8
Ethylenediaminetetraacetic acid (EDTA)	Sigma-Aldrich	Cat#E4884-500G, CAS: 6381-92-6
DNA LoBind 1.5 mL tubes	Eppendorf	Cat#022431021
Trypan Blue	Sigma-Aldrich	Cat#T8154, CAS: 72-57-1
Hemocytometer	Marienfeld	Cat#0640311
DNase I, Bovine Pancreas	Merck/Millipore	Cat#260913, CAS: 9003-98-9

Materials and equipment (optional)

Table 1: Wash medium

Component	Final Concentration	Volume
IMDM	90%	N/A
FBS	10%	N/A

Media can be made beforehand, sterile filtered and stored at 4°C. Prepare 500 mL, which is sufficient to process four batches of cord blood.

Table 2: Freezing medium

Component	Final Concentration	Volume
Fetal bovine serum (FBS)	80%	N/A
DMSO	20%	N/A

Media is made fresh. Prepare 10 mL for freezing 20 vials containing 1 mL of cell solution containing a final concentration of 10% DMSO in FBS.

Table 3: HSPC Complete medium

Component	Final Concentration	Volume
StemSpan SFEM medium		25 mL aliquot
SCF	100 ng/mL	1x 25 µL 100ug/mL aliquot
Flt3-Ligand	100 ng/mL	1x 25 µL 100 ug/mL aliquot
IL-6	20 ng/mL	1x 5 µL 100 ug/mL aliquot
IL-3	10 ng/mL	1x 2.5 µL 100 ug/mL aliquot
TPO	50 ng/mL	1x 12.5 µL 100 ug/mL aliquot
UM729	500 nM	1x 2.5 µL 500 nM aliquot
SR1	750 nM	1x 1.875 µL 750 nM aliquot
Primocin	100 µg/mL	50 µL 50 mg/mL solution

Media can be made fresh for each experiment, do not keep for more than one day at 4 °C. 25 mL of the medium is sufficient to fill an entire 384-wells plate, excluding the outer rim of wells.

Table 4: FACS Buffer

Component	Final Concentration	Volume (µL)
Sterile filtered PBS		
BSA	0,05%	
EDTA	1 mM	

FACS Buffer can be made beforehand, sterile filtered and stored at 4 °C. Prepare 50 mL, which is sufficient for five individual experiments including a concentration range of four samples and control.

Table 5: 2x HSPC staining mixture

Antibody	Dilution	Volume (µL)
BV421-CD34	1:20	5
FITC-Lineage mix (CD3/14/19/20/56)	1:20	5
PE-CD38	1:50	2
APC- CD90	1:200	0.5
PerCP/Cy5.5 - CD45RA	1:20	5
PE/Cy7- CD49f	1:100	1
FITC -CD16	1:100	1
FITC-CD11c	1:20	5
FACS Buffer		25.5

2x HSPC staining mixture can be made fresh for each experiment, keep on ice.

Table 6: HSPC harvesting buffer

Antibody	Final Concentration	Volume (µL)
Sterile buffered PBS	99%	N/A
BSA	1%	N/A

HSPC harvesting buffer can be made beforehand and stored at 4 °C

Table 7: DNA buffer (LowTE buffer)

Component	Final Concentration	Volume (µL)
MilliQ water		N/A
Tris	10 mM	N/A
EDTA	0.1 mM	N/A

LowTE buffer can be made beforehand and stored at 4 °C. Prepare 10 mL, which is sufficient for the isolation of ±200 clones.

Step-by-step method details

1. Preparing CMBCs

Note: If starting from fresh cord blood, start with preparation step 1.1. If working with cryopreserved CBMCs, start with preparation step 1.2.

1. **If starting from fresh cord blood:**
   Timing: 2 hours

   Note: The number of CBMCs harvested from cord blood varies between 3-10 x 10^6 CMBCs/mL. The minimal number of cells for a typical genotoxin assay, containing five genotoxin concentrations and one control condition, is 6 x 10^7 CBMCs, requiring the input of 6 - 20 mL cord blood.

   Note: Make sure the umbilical cord blood is not harvested 48 hours or more in the past – and is preferentially less than 24h old to ensure a maximum output of viable $CD34^+$ cells.

   Note: If not directly using CMBC cells, cells may be frozen in preparatory step 1e, and can later be thawed using step 1.2.

a. Dilute umbilical cord blood 1:1 with an equal volume of wash medium (Table 1).
b. Isolate the mononuclear fraction from the sample using a density gradient separation medium such as Ficoll or Lymphoprep.
c. **NOTE:** For easy layering of diluted cord blood pre-filled density gradient separation tubes with insert easing layering of cord blood, such as LeucoSep or SepMate.
   i. For the gradient separation add 15mL of Lymphoprep to a 50 mL tube, or use a pre-filled density gradient separation tube.
   ii. Tilt the tube at 45 degrees and layer the diluted cord blood dropwise on top of the Lymphoprep layer such that the density gradient separation medium and diluted cord blood are not mixed. (Figure 2A)
   iii. corresponding to 0 and spin at 400 x g for 30 min at room temperature. **CRITICAL:** To keep separation layers intact, set acceleration at a low setting, and decelerate without brakes.
   iv. Use a pasteur pipette or 5 mL pipette to carefully pipette the CBMC layer on top of the ficoll layer (Figure 2B).

CRITICAL: As density gradient separation medium is toxic to cell, avoid pipetting the density gradient separation medium during this step.

d. Wash the CBMC suspension by adding two times with 20 mL wash medium and spin down the cell suspension at 350 x *g* for 10 minutes at room temperature.
e. Resuspend cells in 10 mL wash medium and count cells using 0.4% trypan blue and a hemacytometer.
    i. Determine the average number of live cells in each of the four corner squares (4x4 smaller squares each). Live cells do not take up dye, in comparison to dead cells which are stained by the Trypan blue.
    ii. Multiply by 1 x 10^4 to obtain the number of live cells/mL.

    Note: The recommended number of $CD34^+$ cells for each exposure condition is 1 x 10^5, for which a starting number of approximately 1 x 10^7 CBMCs is required.
f. **Optional:** The surplus of CBMCs may be viably frozen and stored in liquid nitrogen, or if $CD34^+$ cells are not isolated directly afterwards.
    i. Resuspend CBMCs with cold FBS at 10^8 cells/mL by gentle pipetting.
    ii. Add an equal volume of cold freezing medium dropwise to the cell suspension while gently shaking the tube to obtain a final solution of 10% DMSO in FBS.
g. Transfer the cell suspension to 1 mL cryogenic vials (maximum final concentration: 5 x 10^7 cells/ml) and freeze cells immediately in a controlled-rate cell freezing
    (-1°C/minute) container at –80 °C. After 24 hours, store cryogenic vials in a liquid nitrogen tank.

    CRITICAL: Exposure to 10% DMSO is toxic for cells. To obtain a maximum number of viable cells after thawing, minimize the time between DMSO addition and freezing.
h. Pellet cells by spinning down at 350 x *g* for 10 minutes at room temperature.
i. Carefully aspirate supernatant and continue with step two: "Isolating CD34+ cells from CBMCs

2. If starting from cryopreserved CBMCs

Timing: 1 hour

Note: Prewarm the Wash medium at 37 °C before starting.

a. Take the cryogenic vial containing the CBMCs from the liquid nitrogen tank and thaw rapidly in a 37 °C water bath.
b. Transfer cells immediately to a 50 mL conical tube. Rinse vial with 1 mL prewarmed Wash medium, and add this to the cells in the 50 mL tube drop by drop.
   Note: You may pool multiple vials from the same biological sample.
c. Slowly add 20 mL warm Wash medium in a dropwise manner.
d. Pellet cells by spinning down at 350 x *g* for 10 minutes at room temperature.
e. Carefully aspirate supernatant, leave 2-3 mL to resuspend pellet.
f. Gently add 15 mL of Wash medium to the cells while shaking the tube.
g. Pellet cells by spinning down at 350 x *g* for 10 minutes at room temperature.
h. Carefully aspirate supernatant and resuspend cells in the remaining volume of medium. Add up to 10 mL with medium and transfer to a 15 mL conical tube.
i. Count cells using 0.4% trypan blue and a hemacytometer. The recommended number of $CD34^+$ cells for each treatment is 1 x 10^5, for which a starting number of approximately 1 x 10^7 CBMCs is required.
   Note: As cell numbers decrease upon thawing, recovery of 20% to 50% of the original frozen cells is expected, with a cell viability ranging between 75% and 95%.
j. Pellet cells by spinning down at 350 x *g* for 10 minutes at room temperature.
h. Carefully aspirate supernatant and continue with step two: "Isolating CD34+ cells from CBMCs

2. Isolating CD34+ cells from CBMCs

Timing: 2 hours

CD34 is a well-established marker of human HSPCs (Stella et al., 1995). Here, CD34-enrichment is performed using magnetic-activated cell sorting with anti-CD34 magnetic beads using the MACS system from Milteny Biotec. Alternatively, other magnetically based separation systems, such as the EasySep system from STEMCELL Technologies may be used. A frequency of approximately 1% $CD34^+$ CB-HSPCs is expected among CBMCs.

1. Prepare 25 mL HSPC Complete medium using the recipe in Table 3.
2. Perform magnetic isolation of $CD34^+$ cells using MACS system from Milteny Biotec (optimized protocol):

a. Spin down the cells at 350 x *g* for 5 minutes, room temperature and aspirate the supernatant.
b. Resuspend the cells with 10 mL of MACS buffer and transfer the cell suspension to a 15 mL tube.
c. Spin down the cells at 350 x *g* for 5 minutes at room temperature. Aspirate supernatant.
d. Resuspend the cells in 300µL MACS buffer.
e. Add 100 µL of FcR blocking reagent. Vortex the reagents bottle before using.
f. Add 100 µL of CD34+ beads and mix well. Incubate for 30 minutes on ice. In case of a high number of cells mix the vial again after 15 minutes and place it back on ice.
g. Wash cells by adding 5 mL MACS buffer, centrifuge 350 x *g* for 5 minutes and remove supernatant.
h. Assemble the magnet on the holder and position the separation column correctly (the wings on the bottom of the separation column must be visible, see Figure 1B of (Rodríguez et al., 2021)). Also, place a collection tube under the column for the liquid collection.
i. Prepare the MACS column for the separation by adding 3 mL MACS buffer.
j. Resuspend the cell pellet in 500 µL MACS buffer and add the suspension onto the column. Collect flow-through.
k. Wash the column 3 times with 3 mL of MACS buffer, collect flow through.
l. Remove column from the magnet and place it on top of new 15 mL tube.
m. Add 5 mL MACS buffer on top of column and gently plunge the column dropwise.
n. Centrifuge the cell suspension at 350 x *g* for 5 min at room temperature.
o. Aspirate supernatant and resuspend the cell pellet in 1 mL of complete StemSpan.
p. Count cells twice using Trypan blue and a hemacytometer.

3. After flushing out the magnetically labeled $CD34^+$ cells from the MACS column, pellet cells by spinning down at 350 x *g* for 5 minutes at room temperature
4. Carefully remove the supernatant and resuspend in 1 mL HSPC Complete medium (Table 3)
5. Count cells using 0.4% trypan blue and a hemacytometer.
6. Seed isolated $CD34^+$ cells at a density of 2.5 x 10^4 to 3 x 10^4 cells/mL in a 6- or 12-well plate filled with 3 or 2 mL HSPC Complete medium, respectively. Total cell counts for 6 and 12-well plates are 7.5-9 x10^4 and 2.5-3 x 10^4 cells respectively.

CRITICAL: To enable accurate cell counts after exposure, seeding the same number of cells in each well is crucial.

7. Place the plate in a humidified, 37°C, hypoxic (5% O_2) incubator with 5% CO_2 for 24 hours.

3. Exposure of CD34+ cells to genotoxic compound(s)

Timing: 3-8 days

After 24 hours of recovery, $CD34^+$ CB-HSPCs are treated with the compound/exposure of choice. To both ensure clonal outgrowth and ensure sufficient exposure, the compound/exposure is titrated to a level resulting in 40-60% lower cell expansion (approximating IC50) after 3-4 days compared to cells treated with the appropriate solvent control. The compound/exposure is incubated with the cells for 72 hours to allow replication and incorporation of DNA mutations.

Note: In addition to chemical compounds, such as genotoxic drugs, the assay can easily be modified to study the mutagenic effects of other external exposures, such as radiation. This exposure needs to be administered after 24 hours of recovery at step 1. This timepoint is chosen to ensure cells have been recovered, and can undergo division cycles in the next 72 hours, to fix the mutation in the genome. In the example X-ray data indicated by Figure 5 we used a Precision CellRad benchtop irradiator, exposing for 1 hit, 2 Gy at maximum voltage (130 kV) and intensity (5 mA).

1. After 24 hours of incubation, add the compound to the cells in the same solution volume. For the control condition, add the same volume of the dissolvent (PBS/DMSO).
2. Place the plate back in the incubator for 72 hours.
3. After 72 hours, transfer cell suspension to a conical 15 mL tube and pellet cells by spinning down at 350 x *g* for 5 minutes at room temperature.
4. Aspirate supernatant, resuspend in 1 mL FACS buffer (Table 4) and transfer to a 1.5 mL microtube.
5. Count cells using 0.4% trypan blue and a hemacytometer and use the resulting cell counts from the unexposed control to count relative survival for all exposure concentrations.
6. Use these relative survival values to calculate a dose-response curve using the R package 'drc', and extract IC40 and IC60 concentrations. Format the input data, named in this example surv_table, using the following structure:

```
> surv_table
      Concentration_nM  Relative_survival
1     0                 1
2     10                0.448
3     33                0.344

> Dr.m = drm(formula = Relative_survival ~
  Concentration_nM,
                  data = d_r, fct = LL.4(names = c("Hill
                  slope","Min", "Max", "EC50")))
```

Critical: If already one exposure concentration is within the IC40-IC60 interval, this sample can be used to proceed with for single cell CB-HSPC sorting in step 3. If not, repeat the exposure experiment with a higher/lower dose of treatment/exposure.

4. HSPC sort and clonal expansion

Timing: 1 day of preparatory work / 4-5 weeks culture time

After 72 hours of incubation in the presence of an exposure concentration, exposed and control cells are stained using a combination of surface markers to define the CB-HSPC population. These CB-HSPCs are sorted as single cells using fluorescent activated cell sorting (FACS). For a representative example of the gating strategy employed see Figure 3.

1. Prepare FACS buffer (Table 4) and keep on ice.
2. Pellet cells by spinning down at 350 x *g* for 10 minutes at room temperature.
3. Prepare 50 µL 2x HSPC staining mixture according to the recipe in Table 5.
   Note: The total volume is distributed between the conditions to take along for HSPC sorting. As a rule of thumb, for approximately 1 x 10^5 to 1 x 10^6 cells, 10 to 20 µL 2x HSPC staining mixture is needed.
4. Carefully remove the supernatant and resuspend the pellet in an equal volume of FACS buffer and 2x HSPC mixture.
5. Incubate for 1 hour on ice in the dark.
6. Prepare a 384-well-plate with 75 µL HSPC Complete medium (Table 3) per wel, except the outer wells. One half of a 384-well plate is required for one condition (genotoxin concentration). This should result in at least 8 clonal expansions, which is sufficient for downstream sequencing analyses (see introductory note Step 6, Variant calling and filtering).

Note: Fill all outer wells with 75 µL PBS or MilliQ water to prevent evaporation of the HSPC Complete medium in the inner wells.

7. Sort single cell HSPCs with the gates set for single cells $Lin^-CD34^+CD38^-CD45RA^-$ according to the following steps:
   Note: Sorting equipment may vary on facility availability. In all our experiments, we have made use of an Sony SH800S cell sorter using a 100 µM filter chip, sorting at a maximal speed of 2,000 events/second.
   a. First, exclude non-cell events by setting a 'cells' gate including all larger elements (Figure 3A)
   b. Second, sort only single cells by removing all cells with a larger forward-scatter-area (FSC-A) compared to forward-scatter-height (FSC-H) (Figure 3B).
   c. Set a gate for all FITC unmarked Lin- cells (Figure 3C) to exclude differentiated cell types.
   d. Set a gate for CD34+ cells (Figure 3D, vertical bar).
      Optional: A distinction can be made between CD90+ hematopoeitic stem cells (HSCs) and CD90- multipotent progenitors (MPPs), which is indicated by the upper and lower gates.
   e. Gate for CB-HSPCs by setting a gate for CD38- CD45RA- cells (Figure 3E) and sort these cells in a 384-well plate.
8. Wrap the 384 well culture plate (with lid) in transparent polyethylene wrap and directly transfer to a humidified, 37°C, hypoxic (5% O_2) incubator with 5% CO_2.
9. Allow single cells to expand clonally for 4-5 weeks.

Pause point: Clonal expansion till confluency typically takes 4-5 weeks. During this time, plates can be inspected (once/twice a week) to keep track of the clonal outgrowth using a light microscope.

5. Harvesting HSPC clones, DNA extraction and WGS

Timing: 1 day preparatory work, 1 week sequencing

Note: After 4-5 weeks the clones will be ready to harvest. Clones to be harvested should cover at least 50% of the well surface to ensure a sufficient DNA yield for WGS. See example of a fully confluent clonal expansion in Figure 4.

Note: To enable sufficient filtering for germline mutations, sequencing DNA from at least three CB-HSPC clones from the same donor is required. After the mutation calling and filtering steps, vcf files containing somatic variants are created, which can be compared to CB-HSPC exposure data from other donors (section 6)

1. To prevent cells sticking to the pipette tip, which reduce DNA yield, coat the tip of the pipet with cold HSPC harvesting buffer (Table 6) by pipetting HSPC harvesting buffer up and down (Table 6).
2. Fiercely pipet up and down the medium in the well with a pipet set at 75 µl and collect the cell suspension in a labeled 1.5 mL DNA LoBind microtube.
3. Rinse the well with 75 µl of cold HSPC Harvesting buffer.

Note: Make sure to scratch the bottom surface of the well, including the corners. As cells may stick to the bottom of the well, check under the microscope to make sure you have collected all cells.

4. Pellet cells by spinning down at 350 x ***g*** for 5 minutes at 8 °C and remove the supernatant.

Pause point: The pellet may be stored at -20°C for at least 6 months.

5. Isolate DNA using the Qiagen QIAmp DNA Micro kit following the "Purification of genomic DNA from cultured cells using the QIAamp DNA Micro Kit" User-developed protocol with the following adjustments:
   a. Add 2 µl of RNase A after addition of buffer AL, incubate for 2 minutes before adding proteinase K.
   b. Incubate for 30 minutes at 56 °C instead of 10 minutes.
   c. Elute the DNA by loading the column with 50 µL of Low TE buffer (Table 7), reload the eluate on the column and spin again for optimal yield.
   d. Determine the DNA concentration using Qubit using 1 µL eluted DNA for each clone. The expected yield varies between 0.5 to 3 ng/µL. The minimal recommended DNA yield to proceed for whole genome sequencing is 1ng/µL.
6. Use at least 25 ng (50 ng recommended when available) of genomic DNA and construct whole genome libraries using a Truseq Nano kit, using a 500 bp insert size. For full instructions, see the TruSeq® Nano DNA Library Prep Reference Guide by Illumina.

7. Sequence each library on an Illumina Nextseq or Novaseq 2x150 bp to an average coverage depth of 15x.

6. Variant calling and variant filtration

For reliable downstream mutation analyses, a true positive rate for somatic variants of > 90% is required. Our approach implements the GATK best practices for data pre-processing and genotyping using the GATK HaplotypeCaller (NF-IAP), after which variants are extensively filtered using a python-based filtering scripts (SMuRF). Alternative somatic variant calling pipelines and filtering strategies may be used when yielding a sufficient fraction (> 90%) of true somatic variants.

Critical: The number of germline variants present in the genome of a typical CB-HSPC far exceeds the number of somatic mutations, even after genotoxic exposure. To ensure filtering of all germline variants, WGS data from the same CB donor can be used within each NF-IAP and SMuRF run.

Timing: 1 week

1. Perform BWA read mapping, quality control, GATK variant calling, variant filtering and variant annotation using the NF-IAP pipeline.
2. First, configure the run.config file, provided as a template template_run.config. setting the folder containing all FASTQ files at 'fastq_path'.
3. From /path/to/output/donorX/NF_IAP/, run:

```
$ nextflow run /path/to/scripts/NF-IAP/nf-iap.nf \
        $       -c {run.config} \
        $       --out_dir $PWD -profile slurm
```

4. Next, perform further variant filtering and annotation using SMuRF to obtain high-quality somatic mutations. From /path/to/output/donorX/SMuRF/, run:

```
$ . /path/to/scripts/SMuRF/venv_3.6/bin/activate

$ /path/to/scripts/SMuRF/SMuRF.py \
$     -b "/path/to/output/donorX/NF_IAP/data/BAMS/*.bam" \
$     -i/path/to/output/donorX/NF_IAP/data/VCFS/VCF/*.filtered*.vcf" \
$     -c {SMuRF_config.ini}
```

Note: To filter out additional recurrent artefacts, filtering using a mutation blacklist is recommended. See data and code availability for blacklists (GRCh38 reference genome build).

5. Then, filter only mutations that are clonally present in 1 clone and not subclonally present in any other clone. Next, split the vcf in single-sample vcfs.

```
$ grep "^#" {SMURF_filtered.vcf} > {SMURF_filtered_manual.vcf}
$ grep -v "^#" {SMURF_filtered.vcf} | \
$ grep “;CLONAL_SAMPLES=1;” | \
$ grep “;SUBCLONAL_SAMPLE-0” >> {SMURF_filtered_manual.vcf}

$ bash /path/to/scripts/SMuRF/scripts/split_in_single_sample_vcfs.sh  \
     {SMURF_filtered_manual.vcf}
```

6. Run the GRIDSS-LINX pipeline to detect somatic structural variants and copy number variations using .bam files. In contrast to small somatic variants, a direct comparision to a control clone is required. Instructions for running the pipeline using docker or docker via singularity can be retrieved from: https://github.com/hartwigmedical/gridss-purple-linx. From /path/to/output/donorX/SV/ run:

```
$ singularity run docker://gridss/gridss-purple-linx:latest \
      -n /path/to/output/donorX/NF_IAP/data/BAMS/*CONTROL1.bam \
      -t /path/to/output/donorX/NF_IAP/data/BAMS/*TREATMENT1.bam \
      -s “TREATMENT1” \
      --ref_genome_version HG38 \
      --snvvcf /path/to/output/donorX/NF_IAP/data/VCFS/VCF/*.filtered*.vcf
```

Optional: It is possible to add somatic variants to improve the copy number fit, using the option –snvvcf.

7. Analysis of in vitro induced somatic mutations

Note: all analyses described below will be performed in R.

1. Load the data in R, filter on the single base substitutions (SBS).

```
> vcf_files = list.files("/path/to/output/donorX/SMuRF/",
                  pattern = ".*SMURF_filtered_manual_.*.vcf",
                   full.names = TRUE)
> sample_names = gsub(".*_manual_(.*).vcf", "\\1", vcf_files)
> grl = read_vcfs_as_granges(vcf_files, sample_names, ref_genome,
                                type = "all")
> grl_sbs = get_mut_type(grl, type = "snv")
```

2. When the information about the treatment conditions has been added, the lengths of the grl_sbs object can be used to visualize the SBS mutation load and compare it to control conditions to determine accumulation of additional mutations.

```
> treatment = c("CONTROL" "TREATMENT1")
> info_mut_load = data.frame(treatment,
                             sample_names,
                             sbs_load = lengths(grl_sbs))
> ggplot(info_mut_load,
    aes(x = treatment, y = sbs_load, fill = "treatment")) +
    geom_boxplot(outlier_color = NA) +
    geom_jitter(color = 'black')+
    ggpubr::stat_compare_means(comparisons = list(c("CONTROL",
    "TREATMENT1"))
```

Note: In the analysis displayed above, the number of SBS mutations are quantified and compared against control. Similar analyses can be performed for double base substitution (dbs), indel and structural variation (SV) mutation types. For dbs and indels, these can be retrieved by using:

> grl_dbs = get_mut_type(grl, "dbs")

> grl_indel = get_mut_type(grl, "indel")

SV counts can be retrieved from the length of the SV-linx table.

3. Make a mutational matrix with 96-trinucleotide SBS profiles per cell, average per treatment, and subtract the control condition to create a SBS profile per treatment. Finally, plot these profiles. (Figure 5D).

```
> mut_mat = mut_matrix(grl_snv, ref_genome)
> mut_mat_merge = do.call(rbind, lapply(unique(treatment),
                          function(this_treat) {
                           rowSums(mut_mat[ ,treatment == this_treat])
                           }))
                           %>% `colnames<-`(unique(treatment))
>  mut_mat_merge_corr = mut_mat_merge[ ,!grepl("CONTROL",
                    colnames(mut_mat_merge)] - mut_mat_merge[ ,"CONTROL"]

> plot_96_profile(mut_mat_merge_corr)
```

Note: dbs counts can be obtained using get_dbs_context, and can be plotted using plot_dbs_contexts. Similarly,indel counts can be obtained using can similarly count_indel_contexts and plot_indel_context functions (Figure 5B, D). For more details, see the MutationalPatterns vignette at: https://bioconductor.org/packages/release/bioc/vignettes/MutationalPatterns/inst/doc/Introduction_to_MutationalPatterns.html

4. Structural variations can be plotted in a circos format using the linx.vis_sv_data.tsv table returned by the GRIDSS-PURPLE-LINX pipeline. Due to the relatively low sequencing depth, false positive allelic imbalances can be detected, which can be removed from each linx.vis_sv_data.tsv, here loaded as SV_table:

```
> SV_table = read.delim("/GRIDSS_PURPLE_LINX_output/donorX_treatment1_clone1.txt")
> SV_table = SV_table[!grepl("LOW_VAF|INF|Inf|SGL", SV_table$Type),]
> SV_table = SV_table [!grepl("LOW_VAF|INF", SV_table$ResolvedType),]
```

Optional: To validate the true positive rate of structural variants, it is advised to check the SV positions in the original BAM files using the (IGV).

5 . Use the plot_circos function, provided in the supplementary code to generate a circos plot, to plot all true-positive structural variants from a list of sv_tables (Figure 5F):

```
> plot_control = plot_circos(list(SV_table_control1,
                                  SV_table_control2))
> plot_treatment = plot_circos(list(SV_table_treatment1,
                                  SV_table_treatment2))
```

Expected outcomes

Experimental part

The live $CD34^+$ cell output after MACS isolation is estimated to be ±1% of input CBMCs. On day 4 from the start of the protocol, cells are counted, and the IC50 condition is stained and sorted along with the control sample (untreated condition). The frequency of $Lin^-CD34^+CD38^-CD45RA^-$ can vary between samples. A typical result of a HSPC sort at day 4 of culture is indicated in Figure 4. After 4 to 6 weeks of culture, the clonal outgrowth rates of expaned clones can be determined in treated and untreated conditions. As a result of exposure, there can be a difference in clonal outgrowth rates between untreated and treated conditions after single cell HSPC sort. The single-cell sort and the clonal expansion step ensure for a sufficient amount of single-cell derived DNA used for WGS. Yields for DNA extraction depend on the clone size and typically vary between 0.3 ng/µL and 3 ng/µL when eluting in 50 µL lowTE buffer (15-150 ng total DNA yield).

Bioinformatic analyses

After filtering for germline mutations, the expected average number of unique mutations in unexposed control clones is 29 unique SBS mutations (95% confidence interval 24.0-34.0, two tailed t-test, Figure 5A). Samples exposed to mutagenic compounds are expected to contain more mutations. Depending on the average difference in mutation load and variation between exposure between clones, three or more control and exposed clones are required to reach statistical significance (Figure 1). Based on the mechanism of mutagenesis, each exposure can result in the induction of different mutation types, such as SBS, indel, dbs or structural variation mutations. As an example in Figure 5, cord blood cells have been exposed to 2 Gy irradiation, which induces SBS, indel and structural variant mutations (Figure 5A-C). The mutational profile of induced mutations can also be compared between control samples and exposures. Whereas C>T mutations are the most occuring mutation type in unexposed CB-HSPCs, 2 Gy irradiation results in a broader spectrum, including C>A and T>C mutations (Figure 5D). In addition, 2 Gy irradiation results in the induction of deletions of 5bp or longer, and deletions of single T and C nucleotides in T and C homopolymers, respectively (Figure 5E). In addition, larger (> 50 base pairs) structural variants and chromosomal rearrangements are more often observed in 2 Gy exposed clones (Figure 5F).

Quantification and statistical analysis

The mutagenicity of exposures can be tested by comparing these to unexposed control clones, using a two tailed t-test, described in step 7.2: Analysis of *in vitro* induced somatic mutations. This value should be fdr-corrected for multiple testing when analyzing multiple treatments simultaneously. To determine the mutational spectra of exposed cord blood cells, the spectra of mutations can be inspected, as described in step 7.3 and 7.4. In addition, sbs, dbs and indel *in vitro* mutational patterns can be compared to mutational signatures extracted from large cancer datasets using the "cos_sim" or "cos_sim_matrix" functions in the MutationalPatterns package.

Limitations of the protocol

One of the main limitations of the HSPC treatment protocol described here is the variability and availability of cord blood as the primary material. Indeed, we have noticed slight differences in clonal potential between biological samples after identical treatment conditions. Moreover, although $CD34^+$ cell abundance is generally around 1% of total number of cells in whole cord blood, one cannot exclude irregularities when acquiring primary material from donors. A second limitation is the relatively small time window of culture allowed before HSPCs start differentiating, after which they cannot be clonally expanded. This reduces the effective treatment interval to three days of exposure. This three-day exposure does not allow for inducing large number of mutations over the course of multiple weeks, something which can easily be performed using immortalized cell lines (Kucab et al., 2019). In addition, due to the relatively short timeframe to culture CB-HSPCs, there is no room for generating specific mutant backgrounds prior to exposure.

Troubleshooting

Problem 1:
CBMC solution starts forming tiny clumps of cells (up to 1 mm width/height).
Potential solutions:

- Filter cell solution using a 70 µm filter to remove clumps from solution

- To prevent clumping of cells induced by tangling DNA, supplement washing medium with 0.5 mM MgCl2 + 3000 U/ml DNAse I.
- Clumping most frequently occurs in previously frozen samples. As clumping becomes worse in samples with low viability, ensure that all steps in Step 1 during preparation and freezing are followed accordingly, including slow freezing to -80°C before transferring to liquid N_2 storage.

Problem 2:
Low $CD34^+$ CBMC output after MACS separation (< 1 x 10^5 cells)
Potential solutions:

- Ensure sufficient input of CB cells before starting (>50 mL cord blood and < 24 hours old) the experiment and good cell viability (above 60% viable cells) in preparation steps one and two.
- Compare the color tone of the MACS microbeads to those of a new CD34+ kit. If the color is markedly lighter, switch to a new microbead kit.
- Ensure the separation column is placed correctly in the holder (step 2.2.h), and make sure to remove the separation column from the holder during the plunging step.

Problem 3:
IC50 not reached on day 4 (sort day).

Potential solutions:
- Repeat the assay using a different range of compound concentrations (steps 1-2).
- In the case of no or low cell death, check whether the compound is cytotoxic. When the compound is not cytotoxic, a maximum concentration can be based on the half-life concentration reached in patients from literature.

Problem 4:
No clonal outgrowth after four weeks

Potential solutions:
- Check the plate for smaller colonies, which may still expand. If these are present, wait one or two weeks.
- Growth factors, cytokines and StemSpan SFEM II need to be prepared with a single thaw-freeze cycle only. To prevent any additional thaw-freeze cycles, make single use aliquots for all HSPC Complete medium components.

- Ensure the cells are correctly sorted in the 384-well plate by setting up a mock-run with a 384-well plate covered with parafilm to determine if sorting plate is calibrated correctly before every sort.

Problem 5:
Insufficient DNA yield for WGS after DNA isolation (< 25 ng/clone).

Potential solutions:
- Only harvest clones which have reached over 50% confluency in the well (Figure 4). Wells containing differentiated, large cells typically result in a lower DNA yield.
- After harvesting the cells, freeze cell pellets within 1 hour, steps 5.1-5.5.
- Ensure all pipette tips and DNA LoBind microtubes used during harvesting are coated with HSPC harvesting buffer to avoid cells sticking to plastic surfaces.
- Troubleshoot the DNA isolation protocol kit according to manufacturers' instructions.

Problem 6:
Large numbers (< 1000) of somatic mutations in one or more CB-HSPC clones, including unexposed CB-HSPCs after SMuRF filtering.

Potential solution:
- Unexposed cord blood should have relatively low mutation numbers, with a mean of ± 30.8. Clones may not be filtered against germline data from a matching donor, resulting in an artificially high mutation load. Double-check if the data from all clones within SMuRF run are derived from the same donor.
- When performing experiments using cord blood from different individuals, a sample swap may have happened, and the WGS data from the clone needs to be filtered against germline data from another individual.

Article info 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.vanBoxtel@prinsesmaximacentrum.nl).

Materials availability

We generated a new optimized culture medium for HSPCs by including two new compounds (UM729 and Stem Regenin-1, see table 3) to the previously published HSPCs culture media. For reference to the original protocol please refer to Rosendahl Huber, et al. (Rosendahl Huber et al., 2019).

Data and code availability

This protocol makes use of a nextflow-based bioinformatic pipeline to process raw sequencing data. The code for the full mapping and variant calling pipeline can be retrieved from: https://github.com/UMCUGenetics/IAP. The script used for filtering called variants can be retrieved from: https://github.com/ToolsVanBox/SMuRF. Mutational signature analysis can be performed using the R package MutationalPatterns: https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html. Supplementary code for generating the figures and mutation blacklists can be retrieved from https://github.com/ToolsVanBox/Genotoxin_assay. Sequencing data for CB-HSPCs have been deposited in the European Genome–Phenome Archive (https://ega-archive.org); accession number pending.

Acknowledgements

This research was supported by funding from a NWO Vidi grant to R.v.B, no. 016.Vidi.171.023 and Oncode institute. The authors want to thank M.A. Satzl for providing pictures of the CMBC isolation.

Author contributions

A.R.H. and A.L. performed the experiments. A.R.H. performed the bioinformatic analyses. A.L, A.R.H. F.P, E.B, and J.K wrote the manuscript. All authors read, revised, and approved the manuscript.

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.

References

1. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., Islam, S.M.A., Lopez-Bigas, N., Klimczak, L.J., McPherson, J.R., Morganella, S., Sabarinathan, R., Wheeler, D.A., Mustonen, V., Alexandrov, L.B., Bergstrom, E.N., Boot, A., Boutros, P., Chan, K., Covington, K.R., Fujimoto, A., Getz, G., Gordenin, D.A., Haradhvala, N.J., Huang, M.N., Islam, S.M.A., Kazanov, M., Kim, J., Klimczak, L.J., Lawrence, M., Martincorena, I., McPherson, J.R., Morganella, S., Mustonen, V., Nakagawa, H., Tian Ng, A.W., Polak, P., Prokopec, S., Roberts, S.A., Rozen, S.G., Sabarinathan, R., Saini, N., Shibata, T., Shiraishi, Y., Stratton, M.R., Teh, B.T., Vázquez-García, I., Wheeler, D.A., Wu, Y., Yousif, F., Yu, W., Getz, G., Rozen, S.G., Stratton, M.R., 2020. The repertoire of mutational signatures in human cancer. Nature 578, 94–101. https://doi.org/10.1038/s41586-020-1943-3
2. Cameron, D.L., Baber, J., Shale, C., Papenfuss, A.T., Valle-Inclan, J.E., Besselink, N., Cuppen, E., Priestley, P., 2019. Gridss, purple, linx: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. bioRxiv. https://doi.org/10.1101/781013
3. Crawford, J., Dale, D.C., Lyman, G.H., 2004. Chemotherapy-Induced Neutropenia: Risks, Consequences, and New Directions for Its Management. Cancer 100, 228–237. https://doi.org/10.1002/cncr.11882
4. de Kanter, J.K., Peci, F., Bertrums, E., Rosendahl Huber, A., van Leeuwen, A., van Roosmalen, M.J., Manders, F., Verheul, M., Oka, R., Brandsma, A.M., Bierings, M., Belderbos, M., van Boxtel, R., 2021. Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. Cell Stem Cell 28, 1726-1739.e6. https://doi.org/10.1016/j.stem.2021.07.012
5. Helleday, T., Eshtad, S., Nik-Zainal, S., 2014. Mechanisms underlying mutational signatures in human cancers. Nat. Rev. Genet. 15, 585–98. https://doi.org/10.1038/nrg3729
6. Kucab, J.E., Zou, X., Morganella, S., Joel, M., Nanda, A.S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S.P., Arlt, V.M., Phillips, D.H., Nik-Zainal, S., 2019. A Compendium of Mutational Signatures of Environmental Agents. Cell 177, 821-836.e16. https://doi.org/10.1016/j.cell.2019.03.001
7. Rodríguez, A., Filiatrault, J., Flores-Guzmán, P., Mayani, H., Parmar, K., D'Andrea, A.D., 2021. Isolation of human and murine hematopoietic stem cells for DNA damage and DNA repair assays. STAR Protoc. 2, 100846. https://doi.org/10.1016/j.xpro.2021.100846
8. Rosendahl Huber, A., Manders, F., Oka, R., van Boxtel, R., 2019. Characterizing mutational load and clonal composition of human blood. J. Vis. Exp. 2019, e59846. https://doi.org/10.3791/59846
9. Stella, C.C., Cazzola, M., De Fabritiis, P., De Vincentiis, A., Gianni, A.M., Lanza, F., Lauria, F., Lemoli, R.M., Tarella, C., Zanon, P., 1995. CD34-positive cells: biology and clinical relevance. Haematologica 80, 367–87.

Figures and Figure legends

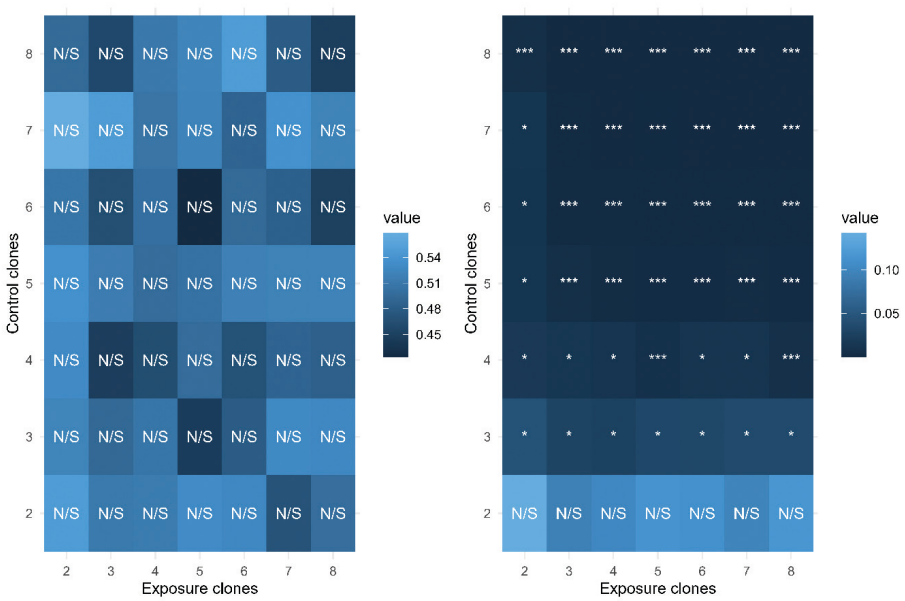


Figure 1: Estimation of number of clones to sequence.
Estimation of number of control and exposed clones to sequence, determined using simulated SBS mutation numbers based on an experiment using control and 2 Gy irradiated CB-HSPCs. Shown are the average *p*-values of 100x simulated results using control and 2 Gy irradiated clones means and standard deviations. N/S = not significant. * = $p < 0.05$, ** = $p < 0.01$.

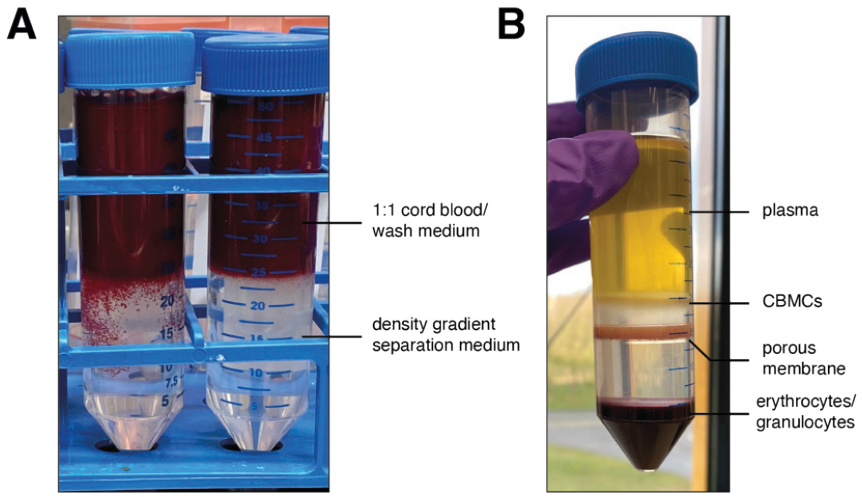


Figure 2: CMBC isolation by density gradient separation
A: 50 mL tubes with the 1:1 Cord blood and wash medium mixture layered on top of the density gradient separation medium. Several minutes after layering, some blood clumps can sink into the density gradient separation medium (left 50 mL tube). This will not affect CMBC yield or quality.
B: Example of succesfull density gradient separation. Different fractions of cord blood are indicated in the figure, with the CMBCs layer located between the plasma and density gradient separation medium. Note: In this example, a 50 mL tube with porous membrane was used.

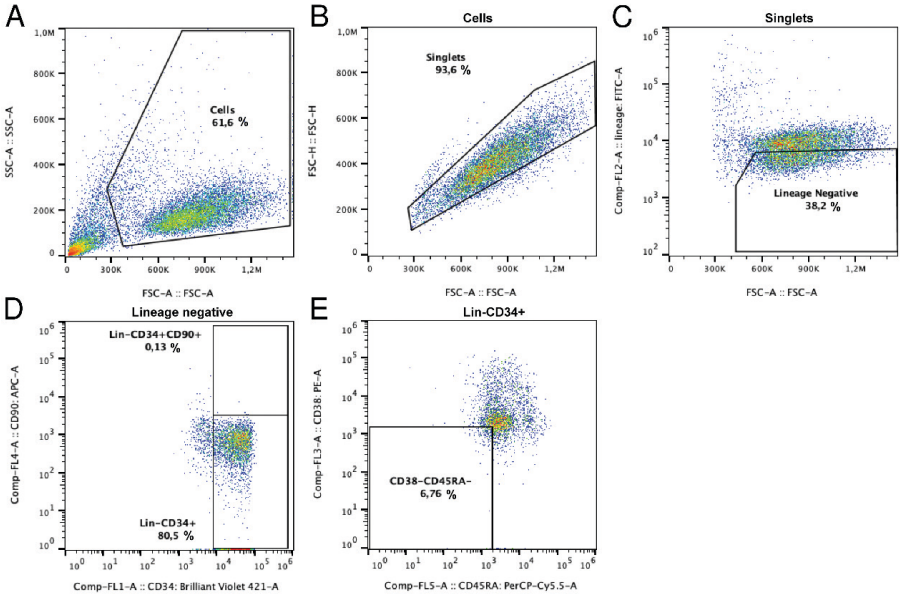


Figure 3: Sorting strategy of CD34+ HSPCs

A: Gate selecting cells based on side-scatter area and forward scatter area.
B: Gate selecting single cells based on forward scatter height and forward scatter area.
C: Lineage-negative cells selection based on lineage-FITC markers (Lin-CD16-CD11c-).
D: Gating of $Lin^{-}CD34^{+}$ cells. Threshold for $CD90^{+}$ is for information only, this treshold is not used to gate and sort cells.
E: Further gate $Lin^{-}CD34^{+}$ for $CD45RA^{-}$ and $CD38^{-}$ status to select for CB-HSPCs. This gate is the gate that is used to single cell sort CB-HSPCs.

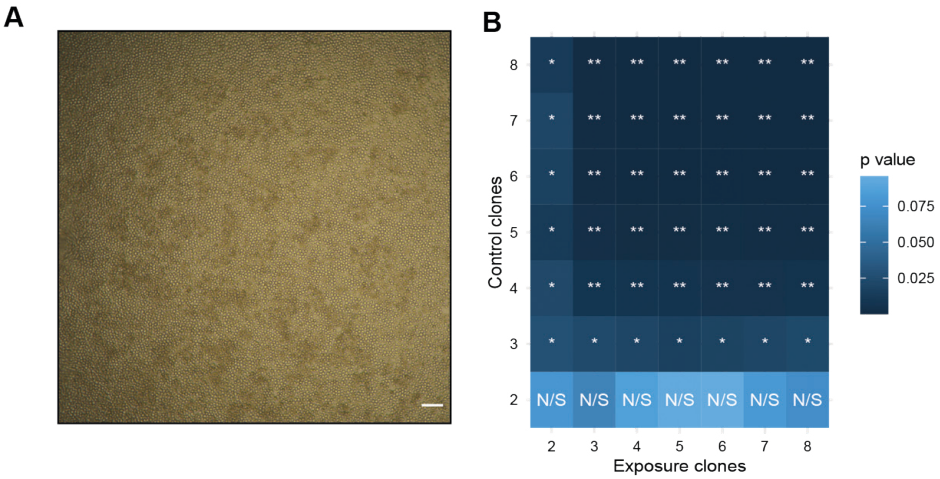


Figure 4: CB-HSPC clonal outgrowth
Center of a single well from a 384 well-plate containing a 100% confluent clonal expansion after 4 weeks. Scale bar indicates 100 µm.

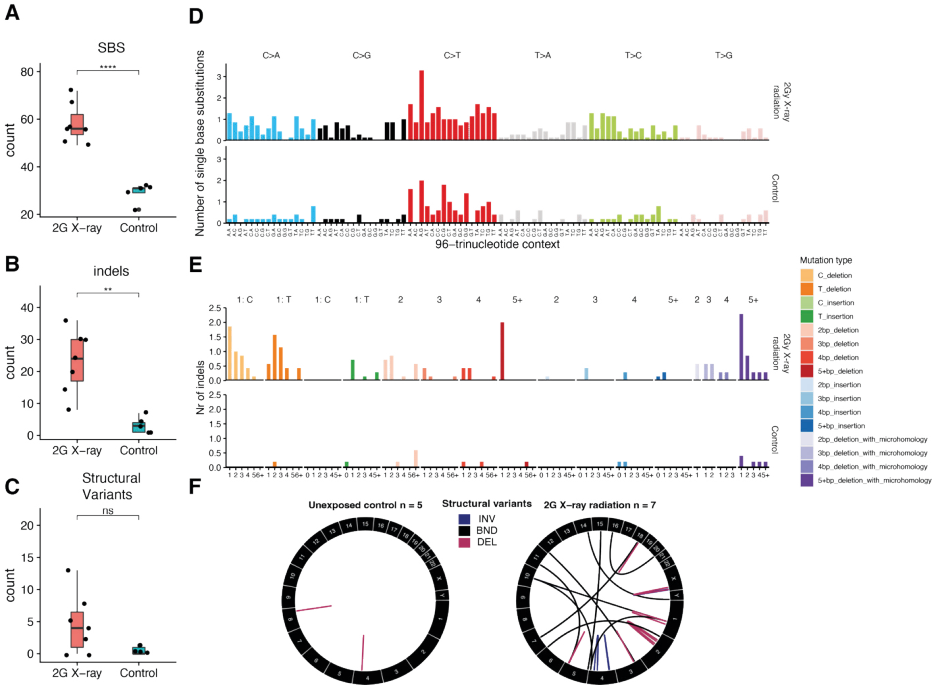


Figure 5: Mutations and mutational patterns in unexposed and exposed CB-HSPCs.
A: Comparison of the number single base substitutions (SBS) present in 2 Gy X-ray exposed cord blood clones and unexposed controls. **** *P*-value < 0.001, two tailed t-test.
B: Similar to A, but for the number of indels. ** *P*-value < 0.01, two tailed t-test.
C: Similar to A, but for the number of structural variants. ns: *P*-value > 0.05, two tailed t-test.
D: The average 96-trinucleotide SBS profiles of 2 Gy X-ray exposed clones and unexposed controls.
E: The average indel context profile of 2G X-ray exposed cord blood clones and unexposed controls, indicating all 83 indel types.
F: Circos plot indicating translocations in an 2G X-ray exposed cord blood clone and an unexposed control. INV = inversions, BND = (inter) chromosomal breakend, DEL = deletion.

Genotoxicity of antiviral nucleoside analog in hematopoietic stem cells

5

Flavia Peci[1,2*], Jurrian de Kanter[1,2*], Niels Groenen[1,2], Laurianne Trabut[1,2], Lucca Derks[1,2], Markus J. van Roosmalen[1,2], Ruben van Boxtel[1,2]

[1]*Princess Máxima Center for Pediatric Oncology, Utrecht, 3584 CS, The Netherlands*
[2]*Oncode Institute, Jaarbeursplein 6, 3521 AL Utrecht, The Netherlands*

* These authors contributed equally

Abstract

Around 90 nucleoside analog (NA) drugs have been approved for clinical use. These NAs are extensively used in the treatment of various diseases, such as cancers and viral infections, including SARS-CoV-2. Due to their mechanism of action, NAs can induce mutations or chain termination during DNA replication not only in infected cells, but also in physiologically normal and uninfected cells. However, the exact mutational consequences of NAs in human cells remain unclear. Recently, we showed that the antiviral NA ganciclovir (GCV), used for the treatment of Cytomegalovirus infection in hematopoietic stem cell transplanted patients, induces mutagenesis in healthy hematopoietic stem and progenitor cells (HSPCs) of pediatric recipients. In addition, we provided evidence that GCV-induced mutagenesis contributes to development of relapses and second malignancies by inducing cancer driver mutations. Here, we systematically screened for the mutagenic consequences of other antiviral NAs using whole genome sequencing (WGS) of *in vitro* exposed HSPCs. Our results showed that most antiviral NA drugs are safe and do not induce enhanced mutagenesis in HSPCs. However, we identified six clinically approved antiviral NAs that were mutagenic and induced a significant number of single base substitutions and double base substitutions. Among these NAs, GCV was found to be the most mutagenic compound.

List of abbreviations

GCV	Ganciclovir
ACV	Aciclovir
PCV	Penciclovir
BVDU	Brivudine
RIBA	Ribavirin
RDV	Remdesivir
AZT	Zidovudine
ABC	Abacavir
MZB	Mizoribine
DDC	Zalcitabine
TFV	Tenofovir
MOV	Molnupiravir
ETV	Entecavir
3TC	Lamivudine
EMT	Emtricitabine
HBV	Hepatitis B
HCV	Hepatitis C

Introduction

The SARS-CoV-2 virus pandemic has taught us that vaccination is the most effective way to prevent and control viral infections[1]. However, vaccines are not available for many viruses, and the success of antiviral treatments largely relies on nucleoside analogs (NAs). The initial NAs took place in the late 1950s, and by the 1960s, they had received approval for clinical use. These NAs continue to be extensively utilized in the treatment of various conditions, including cancer, viral infections, and bacterial infections[2–4]. The success of antiviral NAs lies in their chemical structure and mode of action. The structure created by modifying a naturally occurring nucleotides using standard synthetic methods. Of note, the nucleotide structure provides several sites for potential modifications for drug design purposes, such as changes to the heterocyclic base, the glycosylic bond, the sugar moiety, and alteration of the phosphate group[5]. The structural versatility of NAs provided an opportunity for virologists to intervene and combat the HIV epidemic in the 80s. For example, Azidothymidine (AZT) a thymine analog was approved in 1987 as the first antiretroviral compound for the treatment of HIV+ patients. Despite its toxic side effects, such as bone marrow suppression, as well as development of viral resistance[6], it soon became the first line treatment in clinics. Additionally, antiviral NAs became the gold standard treatment for herpesviruses (HSV) and varicella-zoster virus (VZV)[7,8]. In general, antiviral NAs operate by inhibiting viral DNA/RNA replication machinery in infected cells or by inducing lethal mutagenesis[9]. However, as lethal mutagenicity of virally infected cells is one of the main modes of action of NAs, testing the mutagenicity in healthy cells is essential to ensure safety prior clinical application. Recent work from our group and others has shown that GCV treatment, used in the clinical management of Cytomegalovirus infection, induces mutagenesis in transplanted patients[10,11]. Despite a variety of *in vitro* and *in vivo* tests that have been developed to assess the toxicological profile and safety of NAs, none of them can accurately measure mutagenicity in the whole genome and related genetic risk on a single cell level in a relevant primary human cell type. A variety of *in vitro* and *in vivo* tests have been developed to assess the toxicological profile and safety of NAs. Many of these rely on *in vitro* tests such as the bacterial reverse mutation test (AMES test), mutagenicity tests in selected genes in mammalian cells (HPRT, XPRT, TK kinase), the micronuclei test, chromosomal analysis. Finally, the *in vivo* lethal dose can be assessed in mice. However, one of the limitations of these methods is the single parameter readout. Methods such as the HPRT gene mutation assay provide a readout of mutagenic events in a single gene, instead of performing a more unbiased

genome-wide investigation. Other methods rely on the measurement of drug toxicity based on the metabolic data, namely metabonomic toxicology screening approach[12]. Here, we use a highly sensitive, standardised method to systematically screen antiviral NAs in human HSPCs applying a previously established protocol[13]. Although most antiviral NAs resulted safe to human HSPCs and fail to induce mutagenic events, our screen revealed that 6 compounds are significantly mutagenic and induce low number of additional mutations.

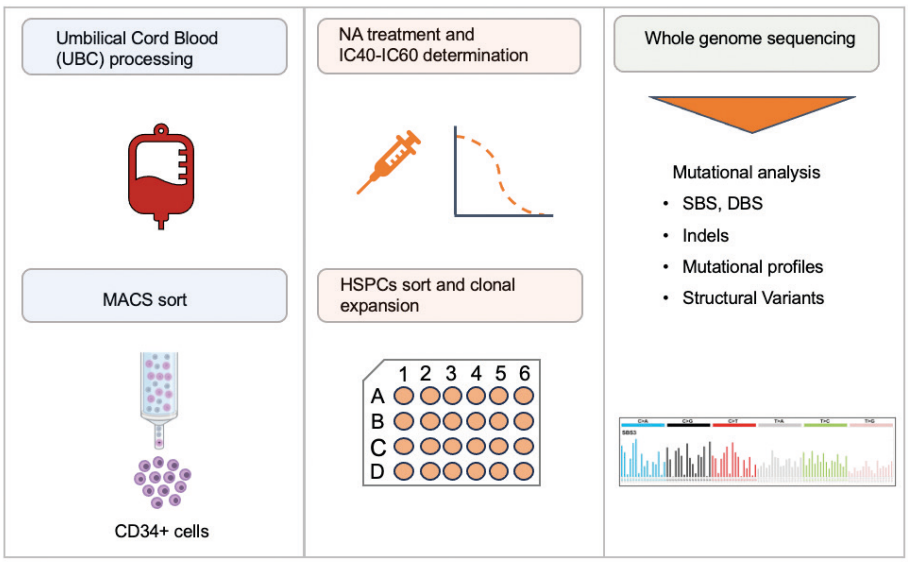
Results

Most antiviral NAs display a safe genotoxic profile in CB-HSPCs.

90 antiviral NAs have been approved by the FDA and EMA (the latter did not approve molnupiravir) for clinical use. From this list, we selected NAs of all four nucleoside subtypes, used in the treatment of different viral infections, such as HIV (zidovudine), SARS-Cov-2 (molnupiravir), Hepatitis B (entecavir, tenofovir, and lamivudine) and Hepatitis C (ribavirin) and Herpesviruses (acyclovir and penciclovir) (see Table 1). To test the mutagenicity of these antiviral NAs, we applied a previous published method[10,13]. Briefly, human HSPCs derived from umbilical cord blood (UCB) were exposed to the compound of interest *in vitro*. After 4 days of exposure, cells that were treated with the IC40-IC60 concentration of the selected compound were clonally expanded to obtain sufficient DNA for whole genome sequencing (WGS). This allowed us to assess all the mutations induced by the compound in the genome of the parental cell. UCB-HSPCs are especially useful for this assay as these cells have a low mutational background and therefore allow for sensitive measurement of even lowly mutagenic compounds. For each NA, initially 3 cells of the same donor were sequenced, except for EMT (two cells) and ACV (four cells from two donors, see Methods). We also applied WGS on 20 untreated control HSPCS from 7 independent UCB donors. Finally, we included previously published WGS data of the highly mutagenic NA ganciclovir as a positive control[10]. For the compounds listed in Table 1, we assessed the following types of DNA damage, single base substitutions (SBS, Figure 1B,1C), small insertions and deletions (indels, Fig. 1D,1E), double base substitutions (DBS, Fig. 1F,1G), structural variants (SVs) and copy number alterations (CNAs). In total, we subjected UCB cells from 21 donors to 15 NAs. From each donor/compound combination we selected 3 HSPC clones for WGS. Clones exposed to 6 out of the 15 tested NAs harboured a significantly increased number of SBS compared to untreated cells (Figures 1B, 1C). These included AZT, MOV), DDC, PCV, BVD and GCV. To confirm these

results, the treatment of AZT, BVD and DDC was repeated with cell of a different UCB donor. The number of mutations were similar between the two repeats for each of these treatments, confirming the robustness of our assay (Fig. 1C, 1E, 1G). Interestingly, while GCV -the most mutagenic compound in our assay- induced an average of 991 mutations per cell, the next most mutagenic compound BVD induced less than one tenth of that (86 mutations). Some other NAs show a low level of mutagenicity, compared to GCV. No SVs or CNAs were detected in any of the treated clones. In addition, no significant increase in the number of indels was detected after treatment with any NA. Finally, only treatment with AZT, DDC, 3TC, and GCV resulted in an increased number of DBS compared to the untreated clones. In conclusion, these results indicate that most of the NAs do not induce any detectable mutations.

A



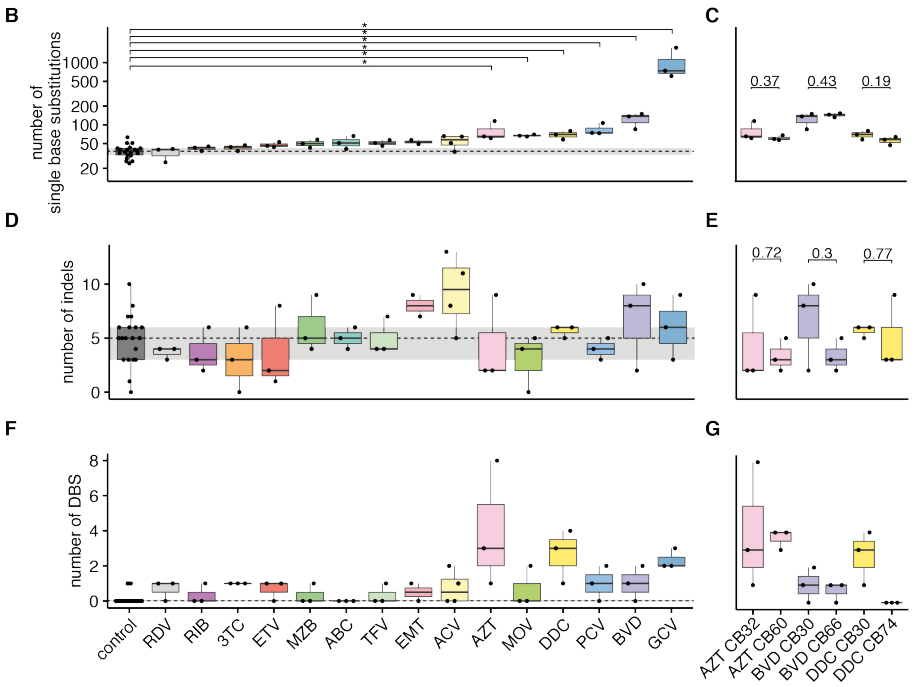


Figure 1. A subset of antiviral nucleoside analogues induces single base substitutions and double base substitutions.

A) Experimental setup of the screening method. B) The number of single base substitutions (SBS) observed in HSPC clones after 4 days of exposure to a variety of nucleoside analogues (NAs), or after no exposure (control). Each dot represents the number of SBS found in a single clone. Each "*" indicates a treatment for which the number of SBS is significantly different from the control clones, calculated by the Wilcoxon test and fdr-corrected. The box plots depict the median (center line), 25th and 75th percentiles (box), and the largest values no more than 1.5* the interquartile range (whiskers); the grey area defines the expected number of SBS according to the age. (C) For the three most mutagenic NAs tested here, the treatment was repeated in cells of a second UBC donor. The number of SBS are shown for the two biological duplicates in boxplots similar to B. P-values are calculated using the Wilcoxon test. (D) The same clones as in B, but the number of indels are shown; the grey area defines the expected number of indels. (E) The same clones as in C, but the number of indels are shown. (F) The same as in B, but the number of double base substitutions (DBS) are shown. P-values were calculated by the Fisher test (taking any value higher than 0 as positive) and fdr-corrected. (G) the same as in C, but the number of DBS are shown.

Mutational profiles of antiviral NAs

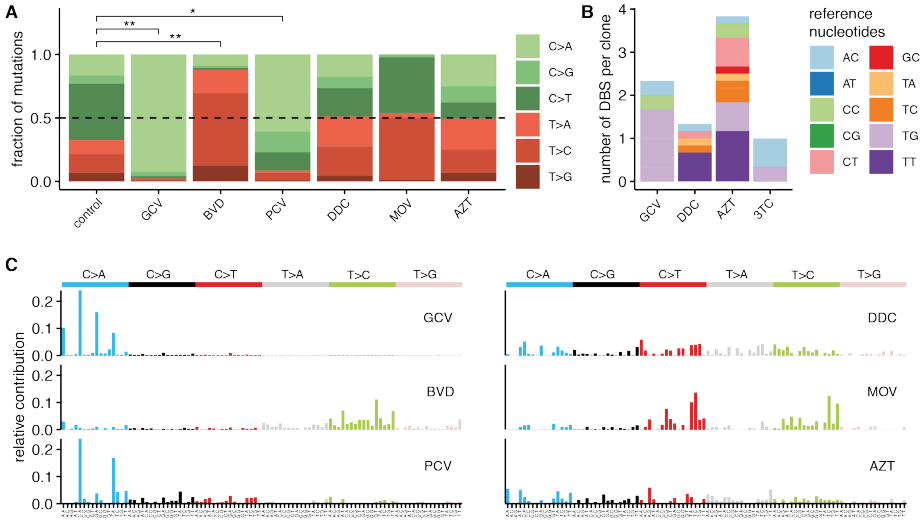


Figure 2. The types of mutations induced differs per NA

A) The fraction of the type of single base substitutions (SBS) induced by each NA. Only NAs are shown that caused a significant number of SBS (Figure 1B). These were corrected for the background mutagenesis, by subtracting the profile of the average control from each treatment profile, before counting the mutation types. P values were calculated by the fisher test and fdr-corrected. * = $p < 0.05$. ** = $p < 0.0001$. (B) The references bases of the double base substitutions (DBS) found in clones exposed to the four compounds that induced DBS: AZT, DDC, GCV, and 3TC. The average number per clone is depicted. (C) The 96 SBS mutation profiles of the treatments shown in A. The profiles of all clones from each donor treated with one NA were averaged. Then, the profiles were corrected for the background mutagenesis *in vitro* by subtracting the average profile of the control clones. The type of mutation that a mutagen induces helps in identifying the mechanism of their mutagenicity. Therefore, we assessed the types of SBS that each of the compounds induced by measuring the mutations in HSPC clones after exposure to the six mutagenic NAs and correcting for the mutations found in the untreated control clones (Fig. 2A). We found that for GCV, PCV and BVD the mutation spectra, after background correction, are significantly different from the background mutations that accumulate *in vitro,* and not merely an increase of normally *in vitro* acquired mutations. Of note, GCV and PCV (both guanine analogues) induced C:G > N:N mutations, while BVD (a thymidine analogue) induces T:A > N:N mutations. In contrast, AZT, MOV, and DDC all induce both types of mutations, hinting towards a distinct mechanism that leads to mutations after exposure to these compounds. In addition, AZT, DDC, 3TC, and GCV seemed to induce a low yet significant number of DBS. In all but 3TC most DBS were TN>NN DBS, suggesting that they are indeed treatment-induced, and not random artifacts which would not show such a specificity (Fig. 2B). Additionally, when also considering the direct sequence context of the SBSs (i.e., the base preceding and following the mutated base), the profile of 96 trinucleotide spectra are unique to each NA compound (Fig. 2C). The 96 SBS profiles show that GCV and PCV induce almost exclusively CA>AA mutations. Exposure to BVD

and MOV also seem to result in mutations in specific contexts, although less specific than GCV and PCV. Finally, DDC and AZT have no distinct mutational profile and thus seem to induce mutations in every possible 3 nucleotide context with roughly the same probability. We have previously reported that GCV-induced mutations display strand asymmetries as well as genomic distribution biases. Mutated cytosines of the C:G reference basepairs are enriched on the leading strand of DNA (and thus the mutated G is enriched on the lagging strand), the transcribed strand of genes, in early replicating regions and in promoters and exons[10]. In addition, GCV-induced mutations display a specific extended sequence context with among others a depletion of A's at the -2 position. Such biases were not found for most of the other NAs, although it is possible that the dataset does not contain enough mutations to reach statistical significance (Figure 3). The only exception is PCV, which is very similar in structure to GCV. Indeed, PCV showed similar replication and transcription stand biases of the C>A mutations, but these were not significant, possibly due to the lower number of mutations the compound induces (Figure 3B,C). Interestingly, PCV is significantly depleted in early replicating regions, while GCV is enriched in these regions (Figure 3D). A possible explanation might be the higher background-to-signal ratio in GCV than PCV, although this should be tested more extensively. In conclusion, these results combined strengthen the idea that for different NAs different mutagenic mechanisms could be involved in inducing the mutations.

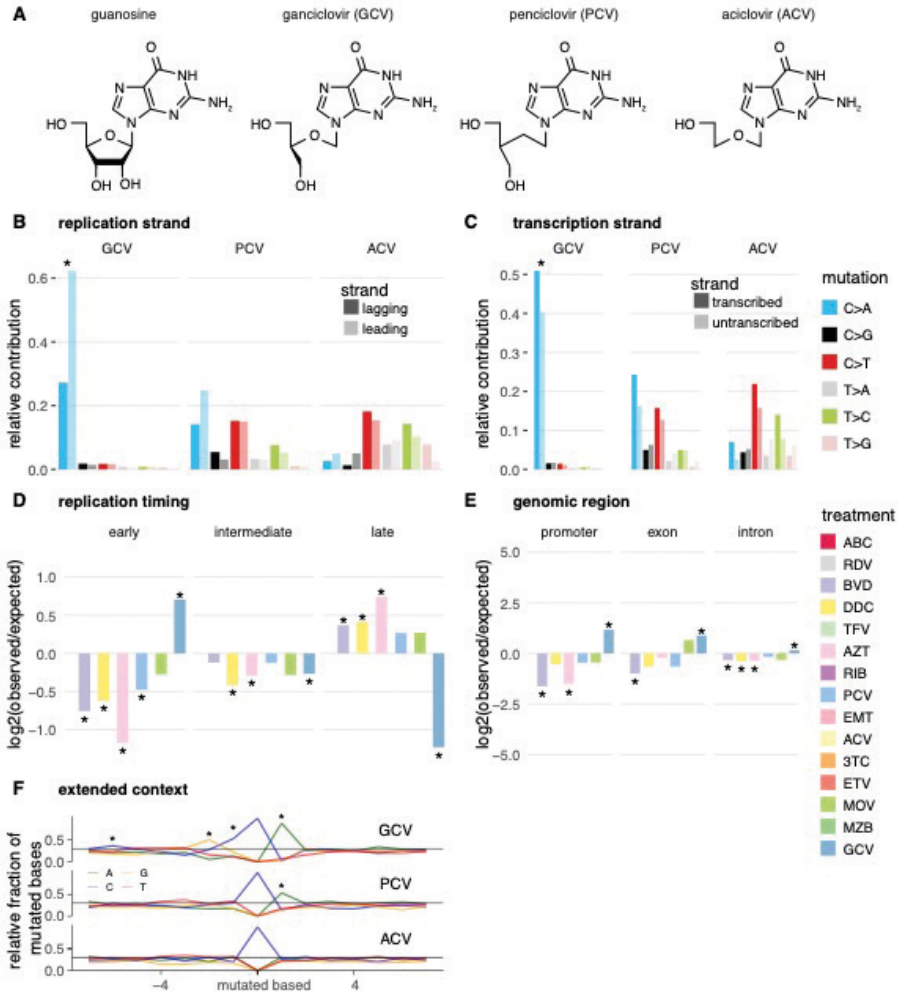


Figure 3. Ganciclovir is the only NA that has strong replication, transcription and genomic location biases

(A) The structures of guanosine and the guanosine analogues ganciclovir, penciclovir and acyclovir. (B) The replication strand (leading or lagging) per mutations type. GCV was the only NA with a fdr-corrected significant bias. In addition, the related penciclovir and acyclovir are shown. (C) Similar to B, but for the transcribed and untranscribed strand of genes. Again, only GCV had an fdr-corrected significant bias. PCV and ACV are also shown. (D) The enrichment or depletion of mutations in regions of early, intermediate, and late replication timing of NAs that induced significant SBSs. GCV was the only NA that was enriched in early and depleted in late replicating regions. (E) The enrichment or depletion of mutations in promoter, exon or intro regions. GCV was the only NA enriched in all three regions. (F) The extended context of mutated bases. GCV and PCV were the only ones with a significant enrichment (fdr-corrected fisher test) of at least context base (mutations were split by C>N and T>N mutations per NA). In addition, ACV is shown.

Antiviral NAs are toxic to mitochondria.

NAs are associated with mitochondrial toxicity[14], as NAs can damage mitochondrial DNA (mtDNA). One example is the development of severe mitochondrial toxicity in HIV+ patients in treatment with antiviral NAs which is an off-target effect and has been documented extensively[15–17]. Although a general mechanism remains to be elucidated, both *in vitro* and *in vivo* tests revealed that DNA (mtDNA) polymerase gamma, Polg, is a sensitive target for inhibition by metabolically active form of Nucleoside Reverse Transcriptase Inhibitors (NRTIs)[18,19]. Here, by using WGS data from sequenced HSPC clones treated with the list of antiviral NAs, we measure the number of mtDNA reads present after treatment. Although most research reports a mtDNA loss upon exposure to reverse transcriptase inhibitor (NRTI)[17,20], we observe a significantly increased number of mtDNA after treatment with TFV, AZT and EMT in HSPCs (Fig. 4A). This could be explained by the recovery time after treatment due to the clonal expansion step in our protocol[13].

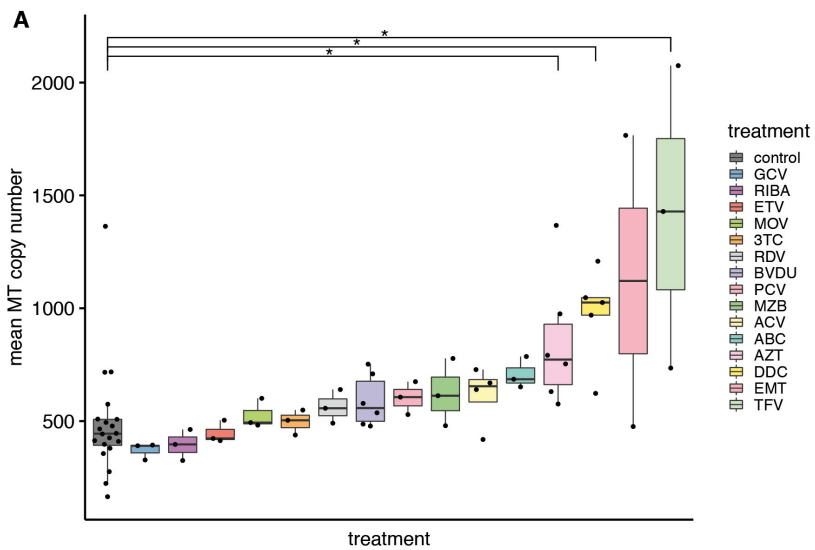


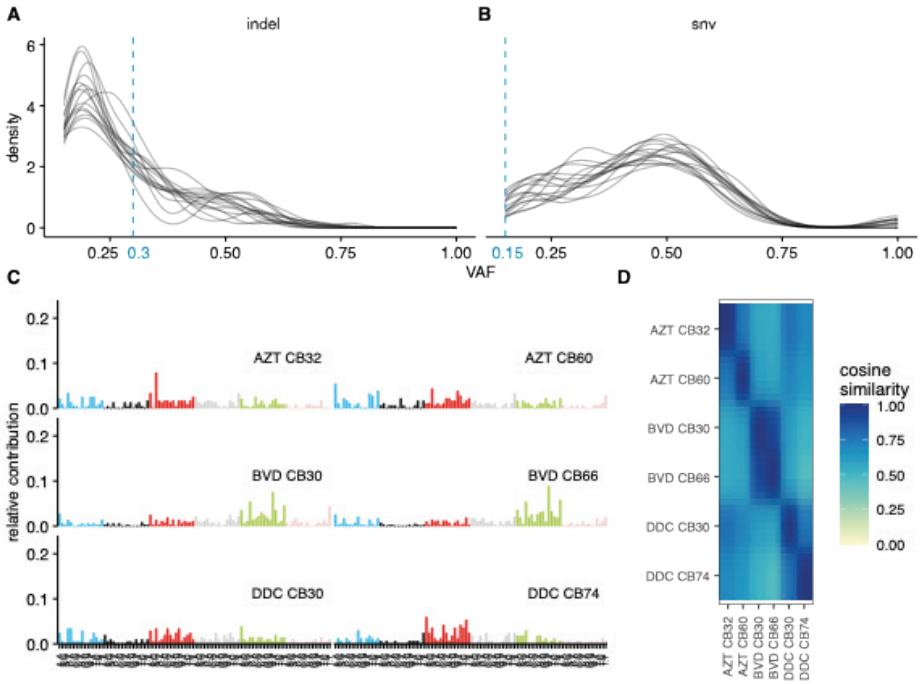
Figure 4. A subset of nucleoside analogues are toxic to mitochondria

(A) The mean of the mitochondrial DNA copy number in HSPC clones, divided per nucleoside analogue. Each dot represents the measurement within a single clone. P-values were calculated by the Wilcoxon test and fdr-corrected.

Table 1

Nucleoside Analogs	Manufacturer	Cat. number	Type of compound	UBC donors	HSPC clones sequenced
Ganciclovir	Sigma-Aldrich	SML2346	Guanosine analog	CB22	3
Aciclovir	Sigma-Aldrich	A0220000	Guanosine analog	CB35, CB44	4
Penciclovir	Sigma-Aldrich	P0035-100MG	Guanosine analog	CB33	6
Brivudine	Fisher scientific	50-194-8398	Thymidine analog	CB30, CB66	6
Ribavirin	Sigma-Aldrich	R9644	Guanosine analog	CB32	3
Remdesivir	Bio-Techne	7226	Adenosine analog	CB30	3
Zidovudine	Selleckchem	S2579	Thymidine analog	CB32, CB60	6
Abacavir	Selleckchem	S5215	Guanosine analog	CB14	3
Mizoribine	Selleckchem	S1384	Guanosine analog	CB47	3
Zalcitabine	Selleckchem	S1719	Cytidine analog	CB30, CB74	6
Tenofovir	Selleckchem	S1401	Guanosine analog	CB31	3
Molnupiravir	Selleckchem	S8969	Cytidine analog	CB44	3
Entecavir	Selleckchem	S5246	Guanosine analog	CB45	3
Lamivudine	Selleckchem	S1706	Cytidine analog	CB45	3
Emtricitabine	Selleckchem	S1704	Cytidine analog	CB39	2

Figure Supplement 1



Supplementary Figure 1. Variant allele frequencies cut-offs and profiles of biological duplicates

(A) The variant allele frequency distributions of indels in untreated and NA-treated HSPC clones. Each line represents the distribution in a single clone. In light blue the cut-off that was used for indel analyses (0.3) is plotted. (B) similar to A, but for single base substitutions. The VAF cut-off here was 0.15. (C) The 96 SBS mutation profiles of independent UCB donors. biological duplicates, meaning HSPC treated clones deriving from two UCB donors per treatment. The profiles of all clones from each donor treated with one NA were averaged. (D) The cosine similarity between the profiles showed in C.

Discussion

Antiviral NAs treatment is primarily administered to immunosuppressed and immunocompromised patients, including those who have undergone organ transplantation and individuals battling cancer. Additionally, patients diagnosed with chronic viral infections like hepatitis are also prescribed antiviral NAs treatment. As antiviral NAs can be mutagenic[10,11,21], it is important to ensure their genetic safety to primary human cells. Here, we systematically investigated the genome-wide mutational profiles of a variety of NAs by exposing primary human HSPCs to a collection of NAs. We selected 16 antiviral NAs approved by the FDA and EMA for clinical treatment of different viruses such as HIV, SARS-Cov-2, HBV, HCV, HSV and applied single cell WGS to HSPCs treated with the NAs. From this list, GCV was the most mutagenic compound in our assay. Recently, the latter was found to be mutagenic in humans[10,11]. Additionally, we included PCV and ACV, which are also guanosine analogues and have a similar structure to GCV (Fig. 3A). As described above, GCV induces more than 10 times the mutations that PCV does, while ACV is not mutagenic in our assay. This observation can potentially be explained by the differences in structures. While GCV is the most similar to guanosine, with only missing one CH-OH group from the ribose ring, PCV also missed the oxygen of the ribose ring and ACV misses another CH-OH group. In theory, it might be possible that GCV is more often incorporated than PCV as its structure is more similar to guanosine and therefore leads to more mutations. Indeed, it has been shown that GCV is incorporated into the DNA of cultured human cells[11]. In contrast, ACV incorporation would completely stop replication as it missed the OH group that is needed for DNA elongation. Of note, available mutagenicity data reports that ACV is primarily a clastogenic compound, only at high concentrations[22]. Nonetheless, we found that antiviral NAs within the same NA group (i.e. guanosine analogs), induce the same type of mutations. Besides, MOV and BVD (thymine analogs) treatment do induce mutations in a specific context, BVD only T>N mutations, but MOV both T>N and C>N mutations. In contrast, DDC and AZT did not show mutations in a specific context. This could be in line with these NAs being incorporated in the human genome during an initial round of replication and resulting in a mutation after mismatching with a nucleotide after the subsequent cell division. A possible explanation could be that these compounds are not incorporated into the DNA of these healthy cells but induce mutations via another, less direct mechanism. A similar observation was recently reported with chemotherapy exposure, where only few compounds induce direct damage, whereas others caused mutations by processes tighten to normal ageing[23]. Alternatively, these NAs

may be incorporated but not in a specific context like GCV, PCV, MOV and BDV seem to be, however this hypothesis still require proper testing. Our data also show that antiviral NAs do not induce SVs, however, aneuploidy induced by AZT and TFV was previously reported in UCB derived T cells of HIV-infected pregnant women treated with AZT and TFV for 4 weeks[24]. Although from one side the use of antiviral NAs has been considered to be adjuvant in the treatment of some type of virus-associated cancers, such as hepatocellular cell carcinoma (HCC) and related long-term complications[25,26], recent studies also highlighted the transplacental genotoxicity of antiretroviral treatments[27,28]. In our *in vitro* setting, the mutagenicity of 6 antiviral NAs is significant compared to control untreated clones, yet the number of mutations acquired after treatment remains low. Based on our work, long-term exposure to the treatment may have important repercussions on the mutagenicity of healthy cells exposed to the compounds and future research is needed to develop genetically safer treatments.

Limitations

As UCB is a primary material collected from donors, variability and availability remain as main limitations to the experimental setup. Furthermore, due to the loss of clonal expansion capacity of HSPCs and increased differentiation capacity over time, the window of treatment exposure is limited to 3 days prior purification for *in vitro* clonal expansion. Additionally, another limitation is the low number of mutations acquired *in vitro*, which influences the mutational pattern analysis.

Conflict of interest

The author declare that the research was conducted in the absence of any commercial of financial relationships that could constitute a potential conflict of interest.

Author contributions

F.P. and J. d. K. wrote the manuscript. F.P. performed the experiments with input from R.v.B. J.d.K. performed the bioinformatic analysis and prepare the figures. All authors approved the manuscript for publication.

Acknowledgments

The authors thank funding bodies and all colleagues for relevant discussion.

Methods

Collection of cord blood samples

Umbilical cord blood samples were collected through the WKZ maternity ward in Utrecht. All samples provided were freshly collected and stored in liquid nitrogen. This study was approved by the Medical Ethical Committee of the Utrecht University Medical Center (protocol number 19-737).

Cell isolation and antiviral nucleoside analog treatment

Mononuclear cells (MNC) fraction was isolated from the cord blood sample using a Leucosep™ tubes (Greiner Bio-One). On the day of the experiment, MNCs were thawed in pre-warmed IMDM media supplied with 10% FBS. Cells were washed two times with IMDM+10% FBS at 350*g* x 10 min at 20°C–25°C and counted by using an automated cell counter (Biorad). Afterwards, CD34+ cells enrichment was performed by using magnetic-activated cell sorting with anti-CD34 magnetic beads from the MACS system (Miltenyi Biotec) according to the manufacturer's instructions. After, CD34+ cells were washed in StemSpan™ SFEM (STEMCELL Technologies) 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) and seeded. Cells were seeded at a density of 0.5×10^5 to 1×10^5 cells/mL in a 12-well plate filled with 2 mL medium, respectively. After 24h of recovery, cells were exposed to the nucleoside antiviral compound of choice at different concentrations and incubated for 72h. For the control condition, the same volume of the dissolvent (PBS/DMSO) was added. After 72h, cells were harvested and span down at 350*g* for 5 min at 20°C-25°C to obtain a pellet. Cell pellets were resuspended in 1 mL FACS buffer and an aliquot was counted with 0.4% trypan blue on a hemacytometer. The resulting cell counts from the unexposed control were used to count relative survival for all exposure concentrations. The microtube corresponding to the IC40-IC60 concentration was sorted as single cells using fluorescent activated cell sorting (FACS) at the SONY Sorter SH800s (SONY).

FACS antibodies and markers

The following antibodies were used in the experimental setting to sort HSPCs: Lineage (CD3/CD14/CD19/CD20/CD56)-FITC (clones UCHT1, HCD14, HIB19, HCD56, 1:20; RRID AB_10644012); 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). The markers used for the cell sorting of HSCs were: Lineage-CD34+CD38-CD45RA-CD90+CD11c-CD16- or Lineage-CD34+CD38-CD45RA-CD49f+CD11c-CD16-. Flow cytometry data were analysed using the Sony SH800S Software (Sony) and FlowJo (BD Biosciences). HSPCs were index-sorted as single cells into flat-bottom 384-well culture plates. Cells were cultured in StemSpan™ SFEM (STEMCELL Technologies) supplied with cytokines. After 3-6 weeks of culture at 37°C and 5% CO2 5% O2, confluent wells were collected for DNA isolation and single cell WGS.

Whole-Genome Sequencing, Read Alignment, mutation calling and filtering

HSPC clones were sequenced with a median genome coverage target of 15x on a Novaseq 6000 (2x150bp). Using Burrows-Wheeler Aligner (bwa) v0.7.17, the sequencing reads were mapped to the GRCh38 reference genome (bwa mem -M -c100). Sambamba v0.6.8 was used for marking duplicates. Mutations calling was performed using GATK. All steps that use GATK were performed with v.4.1.3.0. Variant filtering was done with GATK VariantFiltration using the following filters:

–filter-expression "–filter-expression "QD < 2.0 –filter-expression MQ < 40.0" –filter-expression "FS > 60.0" –filter-expression "HaplotypeScore > 13.0" –filter-expression "MQRankSum <-12.5" –filter-expression "ReadPosRankSum <-8.0" –filter-expression "MQ0 > = 4 && ((MQ0/(1.0 * DP)) > 0.1)" –filter-expression "DP < 5" –filter-expression "QUAL < 30" –filter-expression "QUAL > = 30.0 && QUAL < 50.0" –filter-expression "SOR > 4.0" –filter-name "SNP_LowQualityDepth" –filter-name "SNP_MappingQuality" –filter-name "SNP_StrandBias" –filter-name "SNP_HaplotypeScoreHigh" –filter-name "SNP_MQRankSumLow" –filter-name "SNP_ReadPosRankSumLow" –filter-name "SNP_HardToValidate" –filter-name "SNP_LowCoverage" –filter-name "SNP_VeryLowQual" –filter-name "SNP_LowQual" –filter-name "SNP_SOR" -cluster 3 -window 10

Variant annotation was performed with GATK VariantAnnotator, SNPSiftDbnsfp and SNPEffFilter using the COSMIC v.89, dbNSFP3.2a, and GoNL release 5 databases respectively.

One clone (treated with EMT) was excluded as the median genome coverage was 5x even after two round of sequencing. Two clones from two batches treated with ACV were excluded as their SNP fingerprint did not match to that of the other two (indicating a different donor).

High quality somatic variant were filtered using the in-house produced pipeline SMuRF v2.1.2 (www.github.com/ToolsVanBox/SMuRF). These were mutations that (A) were positioned on autosomal chromosomes, (B) had a GATK phred-scaled quality score R 100, (C) had a mapping quality of 60, (D) had a base coverage of at least 5, (E) had a GATK genotype quality of 99 (heterozygous) or 10 (homozygous), (F) had a variant allele frequency (VAF) of > 0.3 (indels) or >0.15 (SBS), (G) were unique to that clone compared to the other clones within that batch (same cord blood donor and treatment), (H) were not subclonal in any of the clones in that batch.

Were possible, clones were filtered in sets of three to keep the number of background mutations similar in each clone (as including more clones results in a stricter filtering). This was not possible for EMT, where one of the three cells failed QC (average read length, coverage, duplicate reads). In addition, from two different donors, three cells that were treated with ACV were sequenced. From both donors, one of the cells was excluded due to QC, the remaining four were used for analyses.

Mutational profile and signature analysis

The type of mutations (SBS and DBS), the mutational profiles and cosine similarity between mutational profiles were determined using the R package *MutationalPatterns* v3.6.0. This package was also used for determining replication and transcription strand biases, genomic and replication timing enrichment/depletion. SBS had an average variant allele-frequency (VAF) of 0.5 (Supplementary Fig. 1A). SBS with a VAF of 0.15 or lower were filtered out as these could be mutations acquired during the clonal expansion step. Many indels were observed with a VAF between 0.15 and 0.3 in all cases, which also indicates these are not related to the treatment (Supplementary Fig. 1B). Indels

with a VAF of 0.3 or lower were therefore filtered out. All other data visualization was done in R using the *ggplot2* package of the *tidyverse* package suite.

Mitochondrial DNA coverage

The coverage depth of the mitochondrial genome was performed using a previously published pipeline, which was a modification of GATK's Mitochondria pipeline (https://doi.org/10.1016/j.isci.2022.105610). As described, samples with less than 1000x coverage were removed as this indicates potential technical artifacts (n=1).

References

1. Watson, O. J. *et al.* Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. *Lancet Infect Dis* **22**, 1293–1302 (2022).
2. Kataev, V. E. & Garifullin, B. F. Antiviral nucleoside analogs. *Chem Heterocycl Compd (N Y)* **57**, 326–341 (2021).
3. Kadia, T. M. *et al.* An Oral Combination Study of Novel Nucleoside Analogue Sapacitabine and BCL2 Inhibitor Venetoclax to Treat Patients with Relapsed or Refractory AML or MDS. *Blood* **134**, 3926–3926 (2019).
4. Thomson, J. M. & Lamont, I. L. Nucleoside Analogues as Antibacterial Agents. *Front Microbiol* **10**, (2019).
5. Seley-Radtke, K. L. & Yates, M. K. 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 (2018).
6. Richman, D. D. *et al.* The Toxicity of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-Related Complex. *New England Journal of Medicine* **317**, 192–197 (1987).
7. Andrei, G. & Snoeck, R. Advances and Perspectives in the Management of Varicella-Zoster Virus Infections. *Molecules* **26**, 1132 (2021).
8. Birkmann, A. & Zimmermann, H. HSV antivirals – current and future treatment options. *Curr Opin Virol* **18**, 9–13 (2016).
9. Zenchenko, A. A., Drenichev, M. S., Il'icheva, I. A. & Mikhailov, S. N. Antiviral and Antimicrobial Nucleoside Derivatives: Structural Features and Mechanisms of Action. *Mol Biol* **55**, 786–812 (2021).
10. de Kanter, J. K. *et al.* Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. *Cell Stem Cell* **28**, (2021).
11. Fang, H. *et al.* Ganciclovir-induced mutations are present in a diverse spectrum of post-transplant malignancies. *Genome Med* **14**, 124 (2022).
12. Ebbels, T. M. D. *et al.* Prediction and Classification of Drug Toxicity Using Probabilistic Modeling of Temporal Metabolic Data: The Consortium on Metabonomic Toxicology Screening Approach. *J Proteome Res* **6**, 4407–4422 (2007).
13. Rosendahl Huber, A. *et al.* Whole-genome sequencing and mutational analysis of human cord-blood derived stem and progenitor cells. *STAR Protoc* **3**, 101361 (2022).
14. Côté, H. C. F. *et al.* Changes in Mitochondrial DNA as a Marker of Nucleoside Toxicity in HIV-Infected Patients. *New England Journal of Medicine* **346**, 811–820 (2002).
15. Young, M. J. Off-Target Effects of Drugs that Disrupt Human Mitochondrial DNA Maintenance. *Front Mol Biosci* **4**, (2017).
16. Moyle, G. Clinical manifestations and management of antiretroviral nucleoside analog-related mitochondrial toxicity. *Clin Ther* **22**, 911–936 (2000).
17. Maagaard, A. *et al.* Distinct Mechanisms for Mitochondrial DNA Loss in T and B Lymphocytes from HIV-Infected Patients Exposed to Nucleoside Reverse-Transcriptase Inhibitors and Those Naive to Antiretroviral Treatment. *J Infect Dis* **198**, 1474–1481 (2008).
18. Selvaraj, S. *et al.* Antiretroviral Therapy–Induced Mitochondrial Toxicity: Potential Mechanisms Beyond Polymerase-γ Inhibition. *Clin Pharmacol Ther* **96**, 110–120 (2014).
19. Kohler, J. J. & Lewis, W. A brief overview of mechanisms of mitochondrial toxicity from NRTIs. *Environ Mol Mutagen* **48**, 166–172 (2007).

20. Liu, K. ***et al.*** Mitochondrial Toxicity Studied with the PBMC of Children from the Chinese National Pediatric Highly Active Antiretroviral Therapy Cohort. ***PLoS One*** **8**, e57223 (2013).
21. Wutzler, P. & Thust, R. Genetic risks of antiviral nucleoside analogues – a survey. ***Antiviral Res*** **49**, 55–74 (2001).
22. CLIVE, D., TURNER, N., HOZIER, J., BATSON, A. & TUCKERJR, W. Preclinical toxicology studies with acyclovir: Genetic toxicity tests. ***Fundamental and Applied Toxicology*** **3**, 587–602 (1983).
23. Bertrums, E. J. M. ***et al.*** Elevated Mutational Age in Blood of Children Treated for Cancer Contributes to Therapy-Related Myeloid Neoplasms. ***Cancer Discov*** OF1–OF14 (2022) doi:10.1158/2159-8290.CD-22-0120.
24. Vivanti, A. ***et al.*** Comparing genotoxic signatures in cord blood cells from neonates exposed in utero to zidovudine or tenofovir. ***AIDS*** **29**, 1319–1324 (2015).
25. Yin, J. ***et al.*** Effect of Antiviral Treatment With Nucleotide/Nucleoside Analogs on Postoperative Prognosis of Hepatitis B Virus–Related Hepatocellular Carcinoma: A Two-Stage Longitudinal Clinical Study. ***Journal of Clinical Oncology*** **31**, 3647–3655 (2013).
26. Zhang, Q.-Q. ***et al.*** Long-Term Nucleos(t)ide Analogues Therapy for Adults With Chronic Hepatitis B reduces the Risk of Long-Term Complications: a meta-analysis. ***Virol J*** **8**, 72 (2011).
27. Hleyhel, M. ***et al.*** Risk of cancer in children exposed to antiretroviral nucleoside analogues ***in utero*** : The french experience. ***Environ Mol Mutagen*** **60**, 404–409 (2019).
28. Hleyhel, M. ***et al.*** Risk of cancer in children exposed to didanosine in utero. ***AIDS*** **30**, 1245–1256 (2016).

6

Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox

Sjors Middelkamp[1,2], Freek Manders[1,2,3*], **Flavia Peci[1,2,3*]**, Markus J. van Roosmalen[1,2], Diego Montiel González[1,2], Eline J.M. Bertrums[1,2,4], Inge van der Werf[1,2], Lucca L.M. Derks[1,2], Niels M. Groenen[1,2], Mark Verheul[1,2], Laurianne Trabut[1,2], Cayetano Pleguezuelos-Manzano[2,5], Arianne M. Brandsma[1,2], Evangelia Antoniou[6], Dirk Reinhardt[6], Marc Bierings[1], Mirjam E. Belderbos[1], Ruben van Boxtel[1,2,*#]

[1]Princess Máxima Center for Pediatric Oncology, Utrecht, the Netherlands
[2]Oncode Institute, Utrecht, the Netherlands
[3]These authors contributed equally
[4]Department of Pediatric Oncology, Erasmus Medical Center - Sophia Children's Hospital, Rotterdam, the Netherlands
[5]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW) and UMC Utrecht, The Netherlands
[6]Department of Pediatric Hematology and Oncology, University Hospital Essen, Essen, Germany

[*] These authors contributed equally

[#]Corresponding Author.

Cell Genomics, DOI: 10.1016/j.xgen.2023.100389

Summary

Detection of somatic mutations in single cells has been severely hampered by technical limitations of whole genome amplification. Novel technologies including primary template-directed amplification (PTA) significantly improved the accuracy of single-cell whole genome sequencing (WGS), but still generate hundreds of artefacts per amplification reaction. We developed a comprehensive bioinformatic workflow, called the PTA Analysis Toolbox (PTATO), to accurately detect single base substitutions, insertions-deletions (indels) and structural variants in PTA-based WGS data. PTATO includes a machine learning approach and filtering based on recurrency to distinguish PTA-artefacts from true mutations with high sensitivity (up to 90%), outperforming existing bioinformatic approaches. Using PTATO, we demonstrate that hematopoietic stem cells of patients with Fanconi anemia, which cannot be analyzed using regular WGS, have normal somatic single base substitution burdens, but increased numbers of deletions. Our results show that PTATO enables studying somatic mutagenesis in the genomes of single cells with unprecedented sensitivity and accuracy.

Keywords
Single-cell sequencing, whole genome sequencing, primary template-directed amplification, whole genome amplification, somatic mutations, mutational signatures, cancer, Fanconi anemia, hematopoietic stem cells, structural variants

Introduction

Somatic mutations gradually accumulate in each cell during life, which can contribute to the development of age-related diseases, such as cancer[1–3]. Due to the stochastic nature of mutation accumulation, each cell contains a unique set of somatic variants. Amplification of the genome of a single cell is required to obtain sufficient DNA for WGS. One approach for this is to catalogue mutations in clonal structures that exist in tissues *in vivo*[4] or after clonally expanding single cells isolated from tissues *in vitro*[5,6]. However, these approaches can only be applied to cells that have the capacity to clonally expand such as stem cells, precluding analyses of many diseased and/or post-mitotic differentiated cell types[7]. Examples of these are hematopoietic stem and progenitor cells (HSPCs) of patients with Fanconi anemia (FA), who suffer from progressive bone marrow failure and are predisposed to cancer due to an inherited deficiency of DNA repair[8–10]. Much of the research into the mutagenic processes in FA HSPCs has been performed using mouse models[11–13], because primary HSPCs of human patients with FA are difficult to culture and clonally expand *in vitro*[14,15].

An alternative method to clonal expansion is the use of whole genome amplification (WGA) techniques to directly amplify DNA of single cells in enzymatic reactions. However, single-cell WGA technologies have traditionally been hindered by technical limitations due to uneven and erroneous amplification of the genome, leading to artificial mutations, noise in copy number profiles and missing mutations due to allelic dropout[16]. Recently, a novel WGA method, called primary template-directed amplification (PTA), was developed, which contains several critical improvements over the traditionally used multiple displacement amplification (MDA) protocol[17]. Although the amplification biases and allelic dropout rates of PTA are remarkably low, it still generates hundreds to thousands of false positive single base substitutions and indels in each amplification reaction[17,18]. Bioinformatic approaches, such as linked-read analysis (LiRA)[19] and SCAN2[18], have been developed to filter and analyze WGS data of WGA samples. However, these tools still have low detection sensitivities (~10-40%) and therefore most true variants are missed[18,19]. Additionally, while PTA has the potential to enable structural variant (SV) detection in single cells, current tools are not optimized for PTA-based single-cell WGS data.

Here, we developed the PTA Analysis Toolbox (PTATO), which uses a machine learning model to accurately filter artefacts from PTA-based WGS data and

is optimized for SV detection. We demonstrate the applicability of PTATO by analyzing the genomes of normal HSPCs of FA patients and show that, similar to current FA mouse models, these cells have an increased somatic deletion burden.

Results

Training a random forest model to filter PTA artefacts

The artefacts generated by PTA have been shown to follow a specific, non-random 96-trinucleotide mutational profile in WGS data[17,18]. We hypothesized that we could use a machine learning approach to distinguish PTA artefacts from true positive single base substitutions based on multiple genomic features (Figure 1A). For this, we trained a random forest (RF) model, which we previously showed to be highly effective in attributing individual mutations to a specific mutational process[20]. To generate a confident set of true positive somatic single base substitutions for training of the classifier, we sequenced eleven samples of three patients with acute myeloid leukemia (AML) and a clonal lymphoblastoid cell line (AHH-1) using regular bulk WGS as well as single-cell WGS after PTA (Figure 1B, Table S1 and Table S2). Somatic base substitutions that were shared between the bulk and single cell sequenced samples were used as high confidence true variants for training. We combined two approaches to generate a confident set of PTA artefacts for training. First, we PTA-amplified and sequenced the genomes of three single umbilical cord blood-derived HSPCs. Most of the unique somatic variants in these cells will be PTA artefacts, because HSPCs at birth only harbor 20-50 somatic mutations[21–23]. Second, we selected artefacts from the sequenced AML- and cell line PTA samples by implementing and applying a linked read analysis. In this analysis, artefacts are detected because they are not correctly phased with neighboring sequencing reads containing germline variants[19]. The linked read analysis detects a small subset of artefacts with high specificity, but low sensitivity, as only a minority of variants (10-27%) can be linked to an informative germline variant[19]. We varied the ratio between true and false positives in the training set to determine how different ratios affect performance and found that balancing the true and false positives 1:1 yielded the best training results (Figure S1A). In total, 756 PTA artefacts and 756 true positive single base substitutions were used to train the random forest model (Figure 1B).

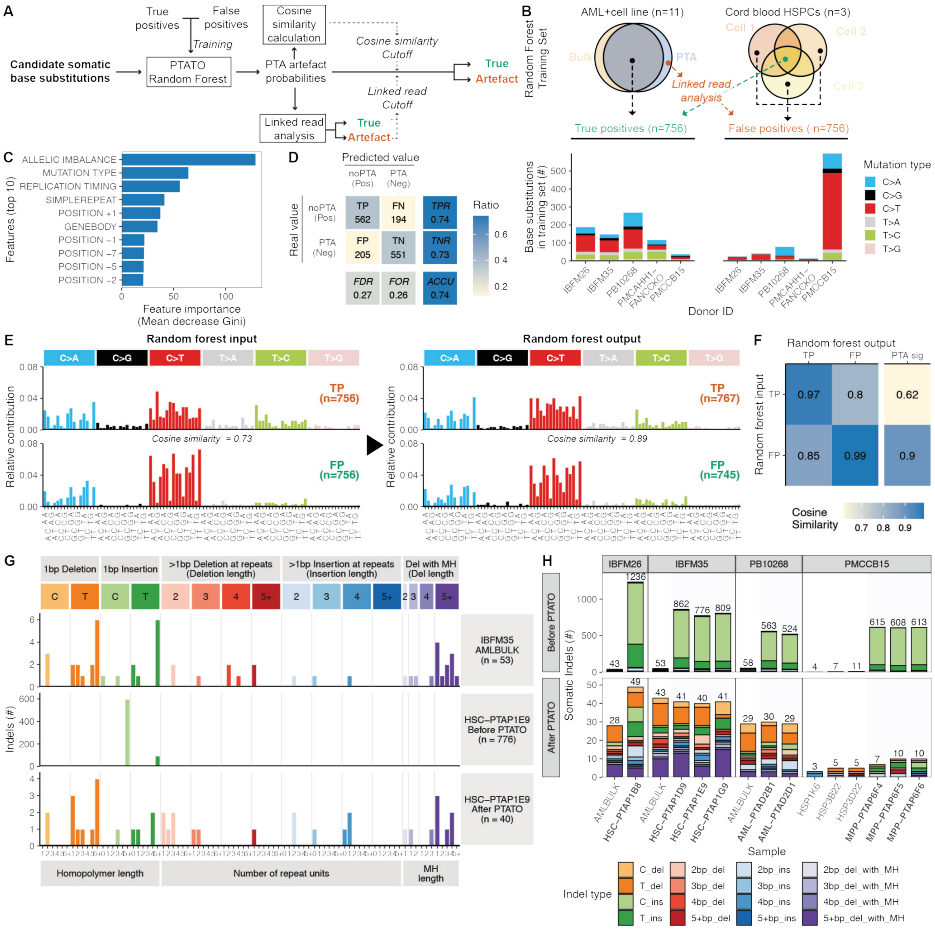


Figure 1: Accurate filtering of PTA artifacts using machine learning and recurrence filtering

(A) Outline of the PTATO workflow to classify candidate base substitutions as true variants or PTA artifacts. The trained PTATO RF model calculates the probability that each variant is a PTA artifact. Subsequently it uses a linked read analysis and cosine similarity calculations to determine a sample-specific probability cutoff.

(B) Overview of the samples and base substitutions that are used as PTA artifacts or true variants to train the RF model.

(C) Importance of the top 10 (out of 26) features used by the RF model to distinguish true variants from PTA artifacts. POSITION indicates the base up- (+) or downstream (-) relative to the mutation.

(D) Confusion matrix visualizing performance metrics of the RF model in classifying out-of-bag variants. TP, true positive; FN, false negative; FP, false positive; TN, true negative; TPR, true positive rate (sensitivity); TNR, true negative rate (specificity); FDR, false discovery rate; FOR, false omission rate; ACCU, accuracy.

(E) The 96-trinucleotide mutational spectra of the base substitutions that were used

as PTA artifact or true-positive input for training the RF model (left) and the profiles of the base substitutions that were classified as true or false by the model during cross-validation (right).
(F) Heatmap showing the cosine similarities between the base substitutions used in the training set and the base substitutions classified during cross-validation and the previously defined mutational signature of PTA artifacts.
(G) Spectra of indels detected in bulk WGS data of AML blasts (top) or before (center) and after (bottom) PTATO filtering of PTA-based WGS data of an HSPC of the same individual.
(H) Numbers and types of indels detected before (top) and after (bottom) PTATO filtering in samples analyzed by bulk WGS or PTA-based WGS (highlighted by blue shading). MH, microhomology; ins, insertion; del, deletion.

To train the RF model, we used a variety of 26 different genomic features, such as the level of allelic imbalance of the region the variant is located in, the mutation type, the 10-basepair (bp) sequence context around the variant, the distance to the nearest gene and replication timing (Figures 1C and S1B). The allelic imbalance is the most important variable in the model (Figures 1C and S1B). This variable is an estimation how well the variant allele frequency (VAF) of a variant matches the modelled pattern of phased VAFs of surrounding germline variants[21]. Other important features for classifying PTA artefacts are the DNA replication timing of the locus the variant is in, whether the variant is in a repeat region and the distance of the variant to the nearest gene (Figures 1C and S1B). These features are likely important, because the distribution of true somatic mutations is known to be biased across the genome, such as depleted in early replicating regions and gene bodies[22]. In contrast, as PTA occurs on naked DNA, PTA artefacts are more randomly distributed over these features in the genome.

The RF model calculates a probability score that a candidate variant is a PTA artefact. As the PTA efficiency and the ratios between true and false positives can vary between samples, a sample-specific cutoff needs to be set above which variants are classified as artefacts. To set an optimal cut-off for each sample, we applied two complementary methods (Figures 1A and S1C-S1G). First, PTATO uses the implemented linked read analysis to classify the small subset of somatic variants that can be linked to informative germline variants as true or false positive. Next, it takes the PTA probability scores for all the variants classified by the linked read analysis and calculates precision-recall curves to determine the optimal cutoff to discriminate these two groups (Figures S1E and S1F). Although this method works well to determine an optimal PTA probability cutoff for most samples, we noted that for some samples accurate precision-recall curves could not be generated because these samples have too few

informative true variants (Figures S1E and S1F). Therefore, we included a second method which calculates mutational spectra at varying PTA probabilities and determines the cosine similarities between these spectra. Subsequently, the cutoff is calculated by hierarchical clustering to separate clusters with similar mutational spectra and low probability scores (containing true variants) from clusters of high probability scores (containing artefacts) (Figure S1G). The RF model was predicted to distinguish artefacts from true positive base substitutions in the out-of-bag sets with an accuracy of 74% (precision = 0.73 and sensitivity = 0.74, Figure 1D) and an area-under-the-curve for precision-recall rates of 0.79 (Figure S1H). Importantly, the 96-trinucleotide mutational spectra of the base substitutions predicted to be false or true variants by PTATO were similar as the profiles of the input PTA artefacts (cosine similarity is 0.99) or true positive variants (cosine similarity is 0.97), respectively (Figures 1E and 1F).

Compared to the base substitution artefacts, the indel artefacts caused by PTA follow an even more specific pattern, which is mainly characterized by C- or T-insertions at long homopolymers (repeats of the same nucleotide) (Figures 1G and 1H)[18]. We found that exclusively filtering indel artefacts that are recurrently called in multiple unrelated individuals and filtering insertions at long (5bp+) homopolymers was even more effective than training a RF model for indel filtering. We created an indel exclusion list containing 5.179.372 indels, which were detected in at least two individuals, across 139 PTA WGS samples of 22 individuals (Figures S2A and S2B). Filtering candidate variants using this list removed most indel artefacts in the samples that were used for training the RF model (Figure S2C), leading to indel burdens and patterns that were comparable (cosine similarity = 0.88) between those found in bulk and PTA-based WGS data (Figures 1G, 1H and S2D). In contrast to SCAN2, which builds a new indel filter list for every analysis if there are sufficient samples[18], PTATO's approach of using a predefined indel filter list is also applicable to small sets of samples and makes indel filtering more comparable between different analyses. Thus, these initial validations demonstrate that PTATO can accurately discriminate true and false positive base substitutions as well as indels using machine learning classification and filtering based on recurrence, respectively.

Validation of the random forest model

We performed several experiments to test the performance of PTATO on samples that were not used in the training set. First, to assess how well PTATO performs on samples containing different ratios of true and false positive base substitutions, we *in silico* mixed different numbers of true base substitutions with a fixed set of PTA artefacts. For this, we collected true somatic base substitutions that were detected in both PTA and bulk WGS samples of two additional AML patients whose samples were not included in the training. Additionally, we obtained PTA artefacts using WGS of an additional PTA amplified umbilical cord blood sample. This *in silico* analysis showed that the performance of PTATO improves with increasing numbers of true variants, especially if there are more than 200 true base substitutions in a sample (Figure S3A,B). Subsequently, to estimate how well PTATO can distinguish true mutations of different mutational backgrounds from PTA artefacts, we *in silico* mutated the trinucleotide sequence context of true positive base substitutions (while keeping the other features the same) to match the 96-trinucleotide spectra of 54 different mutational signatures. This *in silico* mutagenesis experiment revealed that PTATO can accurately detect mutations of the most commonly occurring mutational signatures (e.g., SBS1, SBS5 and SBS18), but also that accuracy is lower for some less prevalent signatures that are very similar to the PTA artefact signature (e.g., SBS30) (Figure S3C).

Secondly, we inactivated the *FANCC* and *MSH2* genes in the human AHH-1 lymphoblastoid cell line using CRISPR/Cas9 gene editing (Figure S4). Inactivation of these genes and their associated DNA repair pathways has been shown to induce various specific base substitution and indel signatures[23–25], enabling us to test the performance of PTATO on a variety of mutational outcomes. We performed several sequential *in vitro* single cell clonal expansion steps (Figures 2A and 2B), followed by bulk WGS of the expanded (sub-) clones, to calculate the mutation rates in these cell lines. Bulk WGS of the subclones showed that the wildtype, FANCC$^{-/-}$ and MSH2$^{-/-}$ AHH1 clones acquire respectively 10.6, 10.5 and 52.6 base substitutions and 1.02, 1.12 and 91.1 indels per day in culture on average (Figures S5A and S6A). Subsequently, after further *in vitro* expansion of the subclones (Figure 2A), we sorted single cells of each subclone and performed WGS after PTA. The standard GATK-based somatic variant calling pipeline (Methods) without PTATO filtering detected a 1.37-1.86 fold higher base substitution rate (Figures 2C, 2D and S5A) and a 12-29 fold higher indel rate (Figures 2E and S6A-S6C) in the PTA-amplified wildtype and

FANCC$^{-/-}$ samples compared to the subclones analyzed by bulk WGS. PTATO removed most excess mutations and the calculated mutation burdens after filtering by PTATO and normalization for the fraction of the genome that was callable (STAR Methods) matched the expected mutation burden (based on extrapolation of the mutation rates determined by bulk WGS of the subclones) with a mean accuracy of 89.5% (Figures 2C-2E, S5A, S5B and S6A-S6C). In comparison, SCAN2[18] reported a mutation burden that was on average 50.4% lower than the expected burden (Figure 2D). Filtering by PTATO also improved the similarity between the mutational profiles of the PTA-amplified samples and the profiles of the corresponding bulk WGS-analyzed subclones (Figures 2F, 2G, S5C-S5G, S6D and S6E). The exact number of PTA artefacts in these PTA samples is not known. Therefore, to estimate the number of PTA artefacts before and after PTATO filtering we performed a bootstrapped mutational signature refit against the mutational profiles of the PTA artefacts and the subclones sequenced with regular WGS. This analysis showed that PTATO improved the precision of base substitution filtering over standard GATK-based somatic variant filtering from 59% to 82%, which is only modestly lower (14.6%) than the 96% precision that SCAN2 showed for these samples (Figure 2H). As shown for the *MSH2*$^{-/-}$ cell sequenced after PTA, PTATO can also accurately remove PTA artefacts from samples with low amplification quality (Figure S6F), although the sensitivity to detect true variants is reduced due to uneven coverage and loss-of-heterozygosity over the genome (Figures S5 and S6).

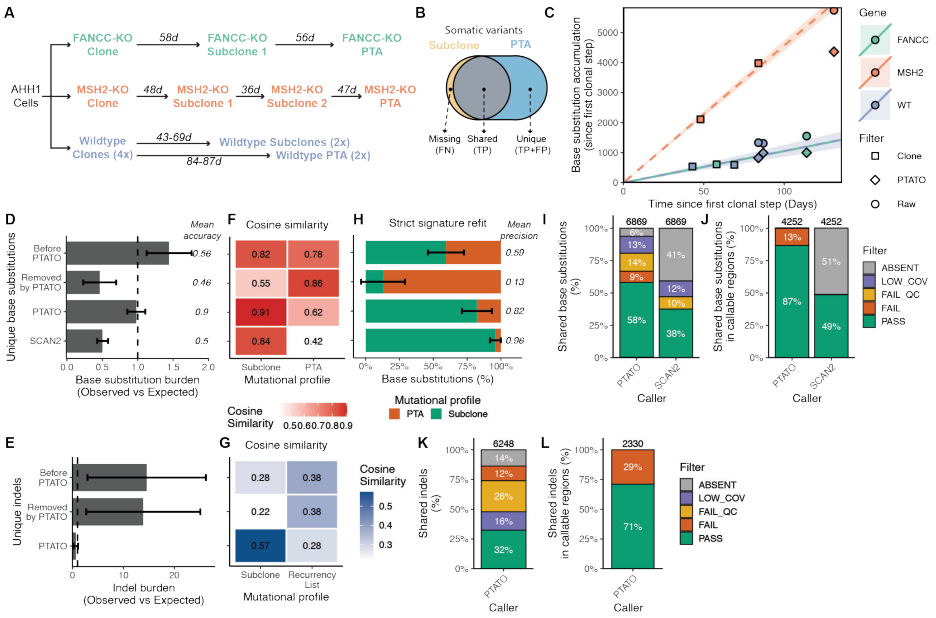


Figure 2: Filtering by PTATO enables accurate analyses of somatic mutation patterns and burdens

(A) Schematic overview of the clonal steps performed for the three types of clonal cell lines generated in this study. Numbers indicate the days (d) in culture between the single-cell sorts, which are used to calculate mutation rates for each cell line.

(B) Venn diagram indicating which variants were used as false negatives (FN), true positives (TP), and false positives (FP).

(C) Accumulation of base substitutions per sample since the first clonal step. The circles and diamonds indicate the number of base substitutions detected in the PTA samples before and after PTATO filtering, respectively.

(D) Observed versus expected number of base substitutions in the PTA samples before PTATO filtering, removed by PTATO, after filtering by PTATO and after filtering by SCAN2. Data are represented as the mean (± SEM) in the four PTA samples.

(E) Observed versus expected (OE) number of indels in the PTA samples before or after filtering by PTATO and after filtering by SCAN2. Data are represented as the mean (± SEM) in the four PTA samples. Accuracy is determined as the mean absolute difference between the OE values and an OE value of 1.

(F) Heatmap showing the mean cosine similarities between the 96-trinucleotide profiles of the unique base substitutions before PTATO filtering, removed by PTATO, after PTATO filtering, or after SCAN2 calling and the profiles of the subclones analyzed by bulk WGS or the previously defined universal PTA artifact signature.[18].

(G) Heatmap showing the mean cosine similarities between the profiles of the unique indels before PTATO filtering, removed by PTATO, or after PTATO filtering and the indel profiles of the subclones analyzed by bulk WGS or the list of recurrent indels used for filtering.

(H) Mean contributions (± SEM) of the universal PTA artifact signature and the mutational signatures of the subclones to the mutational profiles in the four PTA samples before

PTATO filtering, removed by PTATO, after filtering by PTATO, or after filtering by SCAN2. Precision is determined as the mean contribution of the mutational signatures of the subclones to the mutational profiles of the PTA samples.
(I) Fractions of shared base substitutions present in the subclones that are also detected (PASS) in the PTA samples originating from these subclones by PTATO or SCAN2 (SCAN2 could not be used to study indels in these samples).
(J) Fractions of base substitutions after excluding the variants (in both the PTATO and SCAN2 call sets) with low coverage (LOW_COV), low genotype quality (LOW_QC), or undetected variants (ABSENT) as determined by PTATO. Few shared variants are (mis) classified as artifact (FAIL) in the PTA samples.
(K) Fractions of shared indels present in the subclones that are also detected (PASS) in the PTA samples originating from these subclones by PTATO or SCAN2 (SCAN2 could not be used to study indels in these samples).
(L) Fractions of indels after excluding the variants with low coverage (LOW_COV), low genotype quality (LOW_QC), or undetected variants (ABSENT) as determined by PTATO. Some indels are (mis)classified as artifact (FAIL) in the PTA samples (because they are present in the exclusion list or are insertions in long homopolymers).

The somatic variants detected in the (sub)clones should also be present in the corresponding PTA-amplified samples derived from those (sub)clones and thereby form a reliable set of true positive variants. Between 45-69% of the base substitutions (Figure 2I) and 31-56% of the indels (Figure 2K) that were detected in the (sub)clones were also reported in the PTA-amplified cells after PTATO filtering. The clonal variants absent in the PTA-amplified cells were mainly missed due to low coverage and allelic dropout (Figures 2I and 2K), predominately indicating a limitation of the PTA reaction instead of incorrect filtering by PTATO. Importantly, only 10-16% of the base substitutions and 29% of the indels found in both the (sub)clones and the PTA-amplified cells were classified as a PTA artefact by PTATO, showing that PTATO has a mean sensitivity of 86.8% in discriminating detectable true single base substitutions from artefacts in callable loci (Figures 2J and 2L). In contrast, SCAN2 reported on average only 48.8% of these base substitutions shared between these PTA-amplified cells and bulk WGS-analyzed (sub)clones in the callable fractions of the genomes (~78% less than PTATO, Figure 2J). This finding is in line with the ~46% sensitivity reported for this tool[18]. Indels could not be assessed by SCAN2 for these samples, because it required more PTA samples in a single analysis to build a cross-sample filter list. This finding underscores the practicality of PTATO's use of a predefined indel exclusion list instead of creating a novel filter list for each separate analysis.

Thirdly, we further validated the performance of PTATO by applying it to a previously published PTA-based WGS dataset of human umbilical cord blood cells that were treated with a vehicle (VHC) control or with different dosages

of the mutagens D-mannitol (MAN) or N-ethyl-N-nitrosourea (ENU)[17] (Figure S7). We performed strict mutational signature refitting to the universal PTA artefact signature[18] and the SBS1, SBS5 and ENU-associated[26] signatures to estimate respectively the number of false and true positive base substitutions before and after filtering. This analysis showed that filtering by PTATO removed most variants associated with the mutational signature of PTA artefacts with a mean estimated precision of 92%, while keeping most single base substitutions associated with signature SBS5 and/or the ENU-associated signature[26] (Figures S7B-S7E). In the samples treated with a high dose of ENU resulting in a high mutation burden, PTATO detected SBS5- and ENU-associated mutations with an estimated sensitivity of 89% (compared to 60% for SCAN2) (Figures S7D and S7E). The estimated sensitivity to detect true mutations dropped in the VHC-treated control sample with low mutation burden to 37% (compared to 4% for SCAN2) (Figures S7D and S7E). In total, SCAN2 detected 35% less SBS5- and ENU signature-related base substitutions (Figure S7D). Additionally, the 96-trinucleotide profiles detected by SCAN2 in the VHC-samples matched the universal PTA artefact signature with high cosine similarity (0.89 compared to 0.6 for PTATO), suggesting it mostly detected artefacts in these samples (Figures S7C).

Finally, to test how the RF model of PTATO performs on non-hematological samples, we isolated five single cells from a clonal intestinal organoid culture and performed PTA, WGS and PTATO analysis on these cells (Figure S8). Refitting the 96-trinucleotide spectra against the universal PTA artefact signature[18] and a previously described signature of somatic base substitutions accumulating in intestinal organoids *in vitro* (Figure S8C)[6] showed that PTATO can also adequately remove PTA artefacts from single-cell PTA data of intestinal organoids (Figure S8D).

These validations show that PTATO can effectively filter single base substitutions and indel artefacts from PTA-based WGS data from different sources, enabling accurate analyses of somatic mutational burdens, patterns, and signatures in single cells.

Unaltered patterns of indels in most HSPCs of patients with FA

To study the consequences of inactivation of the Fanconi anemia (FA) DNA repair pathway in human HSPCs *in vivo*, we aimed to analyze the genomes of HSPCs of multiple individuals with FA. However, although we flow sorted at least 200 single HSPCs of each of six patients for *in vitro* clonal expansion, only for two patients a limited number of clones (one and eight, respectively) expanded to a size large enough for bulk WGS, underlining the need for direct single-cell WGS. Therefore, we used PTA followed by PTATO analysis to study the genomes of single HSPCs derived from bone marrow aspirates of five different individuals with FA (Table 1). In addition, we analyzed the genomes of bulk AML blasts and three PTA-amplified (pre-)leukemic stem cells from a patient with FA (IBFM35) who developed AML after a failed hematopoietic stem cell transplantation.

Table 1: FA patient characteristics at moment of bone marrow puncture

Individual	Age (years)	Affected Fanconi anemia gene	Fanconi anemia driver mutations	HSC clones	Bone marrow cellularity	Hematological status	Cytogenetic aberrations
PMCFANC01a	7.9-8.4	*FANCC*	c.67delG;c.67delG	1	Moderate/Low	Normal/Mild cytopenia	None
PMCFANC02	15.9	*FANCD1/BRCA2*	c.5213_5216delCTTA; c.9302T>G	8	Moderate	Normal	None
PMCFANC03	15	*FANCA*	c.1361_1370delCCTCCTTTGG; c.1361_1370delCCTCCTTTGG	0	Low	Mild cytopenia	None
PMCFANC06	17	*FANCA*	c.67delG;c.67delG	0	Moderate	Normal	None
PMCFANC08	10.3	*FANCA*	c.2151+1dup;c.2121delC	0	Moderate	Mild cytopenia	None
IBFM35	14.8	*FANCA*	c.3639delT; c.3639delT	0	ND	AML	NA

[a]Bone marrow aspirates from PMCFANC01 were collected at two different time points. HSC, hematopoietic stem cell.

First, we compared the PTATO-filtered base substitutions detected in the HSPCs of individuals with FA with previously generated WGS data of 34 clonally expanded HSPCs of 11 healthy donors[27,28]. This comparison showed that most of the FA HSPCs had similar somatic single base substitution burdens (Figures 3A, 3B, S9A and S9B), patterns (Figures 3C and 3D) and signatures (Figures 3E and 3F) as HSPCs of healthy individuals. Patient PMCFANC02, whose FA was caused by biallelic germline variants in the *FANCD1*/*BRCA2* gene, and AML patient

IBFM35 formed exceptions with respectively threefold and twofold higher somatic base substitution burden than expected for their age (Figures 3A and 3B). The elevated mutation burden in PMCFANC02 is mostly caused by base substitutions characterized by mutational signature SBS3, which is associated with homologous recombination deficiency[29,30], and which is barely detected in the other FA patients (Figures 3E and 3F).

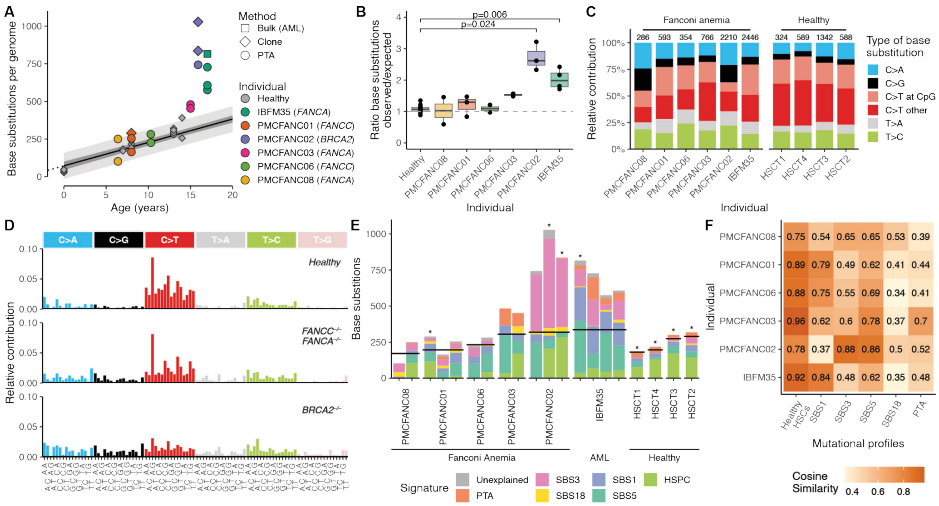


Figure 3: PTATO detects normal single base substitution burdens in most human FA HSPCs

(A) Correlation of the number of somatic single base substitutions per HSPC genome of healthy donors (gray points) and patients with FA. Linear mixed modeling showed that healthy HSPCs accumulate base substitutions in a linear fashion with age.[27,28] The 95% confidence interval and the prediction interval of the model are indicated by the dark gray and light gray shading, respectively.

(B) Ratios between the observed and expected number of base substitutions per genome (sorted on age) based on extrapolation of the age linear mixed model. To match the ages of the patients with FA, only 12 HSPCs of four healthy donors (HSCT1–4, ages 7 to 14) are included in this and following panels. Adjusted p values indicate multiple testing corrected significant differences ($p_{adj} < 0.05$) between three FA patients and the age-matched healthy donors (Bonferroni-corrected Wilcoxon Mann-Whitney test).

(C) Mutation spectra showing the relative contribution of each base substitution type in the genomes of the donors. Numbers above the bar indicate the total number of base substitutions found in the samples from each individual.

(D) The averaged 96-trinucleotide mutational profiles of the HSPCs of the four healthy individuals (HSCT1–4), the patients with mutations in ***FANCA*** or ***FANCC*** (PMCFANC01, PMCFANC03, PMCFANC06, PMCFANC08), and the patient with mutations in ***BRCA2*** (PMCFANC02).

(E) Contribution of base substitution mutational signatures commonly found in blood cells[27,28] to each FA sample or healthy individual (averaged). Horizontal black lines indicate the expected number of base substitutions based on age. Non-PTA samples sequenced with bulk WGS are indicated by an asterisk. For donors HSCT1 to HSCT4, the mean contributions over all samples per donor is shown.

(F) Cosine similarities between the mean 96-trinucleotide mutational profiles of the HSPCs of FA patients with the profiles of the healthy HSPCs from the four age-matched donors and the mutational signatures.

Subsequently, we compared the somatic indel accumulation between HSPCs of patients with FA and healthy bone marrow donors. Only patients PMCFANC02 (*FANCD1*/*BRCA2*) and IBFM35 (*FANCA* and AML) had a significantly increased indel burden compared to healthy HSPCs (also in their bulk-sequenced clones and leukemic blasts) (Figures 4A, 4B and S9C). The relatively high indel burdens in the HSPCs of these two patients did not seem to be caused by a specific type of indel (Figures 4C and 4D). These findings, which are in line with observations in FA mouse models[12] and FA cell lines[23], confirm that PTATO-based filtering of PTA-based WGS data can be used to accurately study somatic mutations in single cells that cannot be clonally expanded *in vitro*.

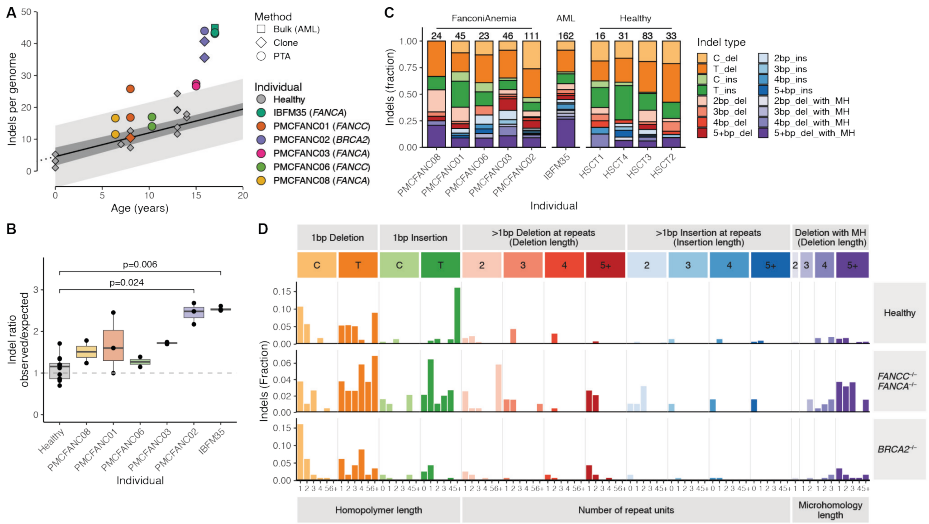


Figure 4: Small insertions and deletions in HSPCs of patients with FA

(A) Correlation of the number of somatic indels per HSPC genome of healthy donors (gray points) and patients with FA. Linear mixed modeling showed that healthy HSPCs accumulate indels in a linear fashion with age.[27,28]The 95% confidence interval and the prediction interval of the model are indicated by the dark gray and light gray shading, respectively.

(B) Ratios between the observed and expected number of indels per genome (sorted on age) based on extrapolation of the age linear mixed model. To match the ages of the patients with FA, only 12 HSPCs of four healthy donors (HSCT1–4, ages 7 to 14) are included in this and following panels. p values indicate multiple testing corrected significant differences ($p_{adj} < 0.05$) between two of the FA patients and the age-matched healthy donors (Bonferroni-corrected Wilcoxon Mann-Whitney test).

(C) Indel spectra showing the relative contribution of the main indel types in the genomes of the donors. Numbers above the bar indicate the total number of indels found in the samples from each individual (without extrapolation for callable loci).

(D) Total averaged indel profiles of the HSPCs of the four healthy individuals (HSCT1–4), the patients with mutations in *FANCA* or *FANCC* (PMCFANC01, PMCFANC03, PMCFANC06, PMCFANC08), and the patient with mutations in *BRCA2* (PMCFANC02).

Accurate detection of structural variants in PTA-based sequencing data

It has been shown that HSPCs of FA mouse models[12] and leukemias[31] and squamous cell carcinomas[32] of human patients with FA have high burdens of somatic structural variants (SVs). Existing bioinformatic tools for single-cell WGS are usually limited to the detection of copy number changes based on read depth[33] and we found that more comprehensive SV calling pipelines for bulk WGS data detect many false positive variants in PTA-based data (Figures 5A and 5B). To study somatic SVs in the HSPCs of the patients with FA, we needed to optimize an SV calling and filtering approach specifically designed for PTA-based WGS data. PTATO integrates calling of SVs by GRIDSS[34] and COBALT[35] based on read depth, B-allele frequencies, split reads and discordant read pairs followed by various normalization and filtering steps tailored for PTA-based WGS data (Figures 5C, S10 and S11).

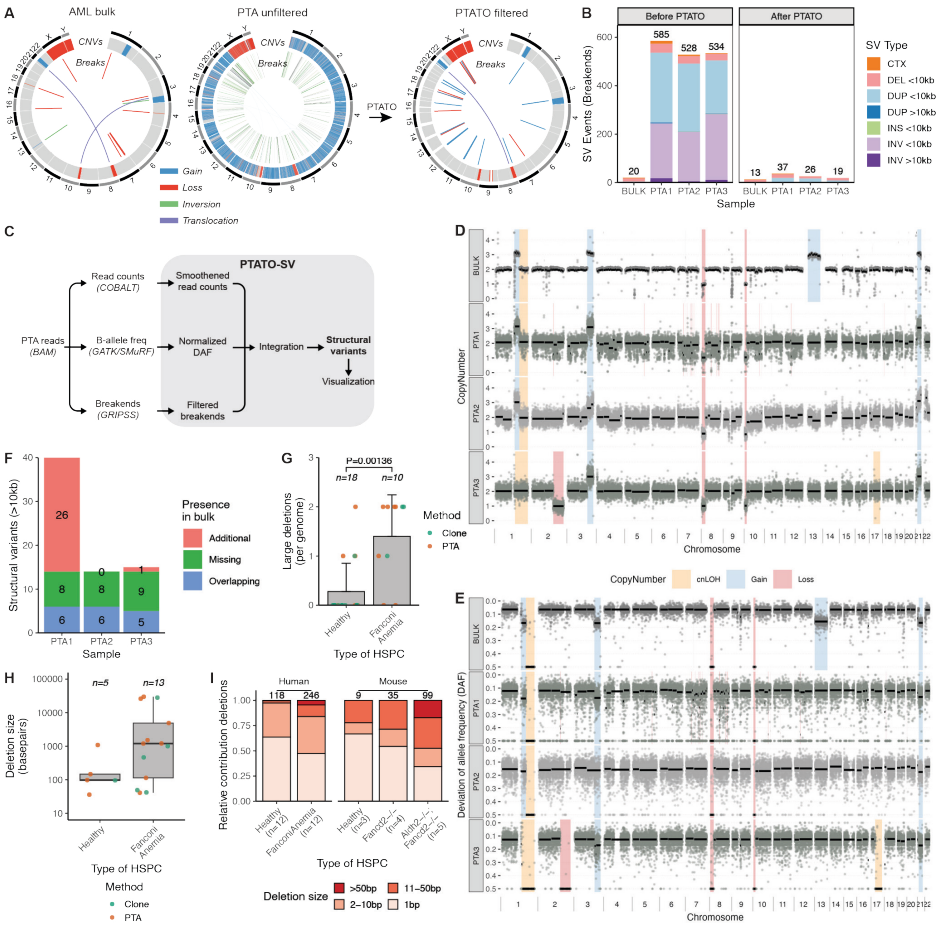


Figure 5: SV filtering by PTATO reveals an increased deletion burden in HSPCs of patients with FA

(A) Circos plots showing copy number variants (CNVs) and balanced SVs in a PTA (left/center) and bulk WGS sample (right) of patient IBFM35. The standard SV calling pipeline for bulk WGS generates hundreds of false-positive calls in PTA samples (left), most of which are removed by PTATO filtering (center), leading to similar SV profiles as a sample sequenced by bulk WGS (right panel).

(B) Number of SV events detected by GRIDSS without filtering by PTATO (left) and the number of SVs remaining after filtering by PTATO (right) in bulk and PTA-based WGS samples of IBFM35.

(C) Schematic overview of the SV calling and filtering strategy tailored for PTA-based WGS data implemented in the PTATO pipeline.

(D) Copy number profiles (100-kb windows) of the AML bulk sample analyzed by the bulk WGS SV calling pipeline and three PTA samples analyzed by PTATO. Background shadings indicate the final copy number call made by PTATO (for PTA samples) or PURPLE (for the bulk WGS sample).

(E) Deviation of allele frequency (DAF) plots (100-kb windows) of the AML bulk sample and three PTA samples. The DAF depicts the absolute difference between 0.5 (perfect heterozygosity) and the actual allele frequency of a germline variant.
(F) Number of SVs (>10 kb in size) that are present in the HSPCs and present ("Overlapping") or absent ("Additional") in the AML bulk or present in the bulk but absent in the HSPCs ("Missing").
(G) Number of deletions (>25 bp) detected by GRIDSS and PTATO in genomes of HSPCs of FA patients or healthy donors (including five cord blood samples sequenced after PTA). Numbers shown above the bars indicate the number of individuals per group. The p value was calculated by Wilcoxon Mann-Whitney test.
(H) Size (in bp) of each detected deletion in HSPCs of healthy donors and patients with FA (no significant difference Wilcoxon Mann-Whitney test). Numbers above the boxes indicate the total number of deletions per group.
(I) Distribution of the sizes of small (detected by GATK for the human samples) and large (detected by GRIDSS for the human samples) deletions in human and mice[12] HSPCs with different genetic backgrounds. The numbers above the bars indicate the total number of deletions analyzed per group.

First, copy number variant (CNV) calling by PTATO started with calculating the coverage over the genome by counting and normalizing the number of reads per kilobase window (Figures S10C and S10D). We noted that the local fluctuations in coverage profiles are recurrent between PTA samples (Figures S10A and S10B). Therefore, we collected copy number profiles of 12 copy number neutral PTA samples and created a panel-of-normals (PON) to smoothen the coverage in test samples (Figure S10E). Subsequently, the smoothened read counts were binned in 100kb windows (Figure S10F) and consecutive windows with similar copy numbers were merged into larger segments (Figures S10G, S10I and S10J). To determine if a segment is a copy number gain or loss, PTATO determined if the distribution of the read counts within the segment is significantly divergent from 1) the distribution of read counts of other segments of the same sample and 2) the distribution of read counts of the same segment in the PON samples (Figure S10H).

Second, the ability to accurately detect germline base substitutions in PTA-based WGS data enabled PTATO to use the B-allele frequencies of germline variants to complement CNV calling (Figures S11A and S11B). PTATO minimized the noise in allele frequencies of germline variants by binning multiple germline variants in 100kb windows, followed by segmentation of the bins (Figures S11A and S11B). Finally, these segments based on allele frequencies were overlapped with the segments based on read depth to determine if segments are copy number losses, gains or copy number neutral loss-of-heterozygosity (cnLOH) regions (Figures 5C, 5D, S11A and S11B).

Finally, the relatively even coverage over the genome in PTA data enabled the detection of split reads and discordant read pairs (break-ends). Hundreds to thousands of artificial SVs, mainly small events that were called as inversions or duplications, were detected in PTA samples by the standard SV calling pipeline (Figures 5A and 5B). PTATO filtered these raw calls using a recurrency list, by excluding SV calls with only one breakpoint junction and by excluding inversion calls that are less than 1kb in size.

We applied the SV filtering to PTA-based WGS data of three HPSCs of a patient with AML (IBFM35) to compare the SV calls in these cells with the SVs detected in the bulk AML sample of this patient. PTATO removed most excess SV calls (Figures 5A and 5B) and determined accurate copy number profiles for these samples (Figures 5D and 5E). Not all SVs present in the AML bulk sample were detected in the PTA samples (Figure 5H). Some SVs (such as the t(3;10) translocation) were missing in the PTA samples due to low coverage around the breakpoints or due to imbalanced amplification (Figure 5A). However, several SVs (such as the gain of chromosome 13) were not detected in any of the single HSPCs despite proper amplification and coverage of these regions, suggesting that these HSPCs are non- or pre-leukemic cells (Figures 5D-5F). To further test PTATO's SV pipeline, we applied it to PTA-based WGS data of a HSPC of AML patient IBFM26 and two single AHH-1 cells. Also in these cells, PTATO generated copy number profiles that were similar as those obtained after bulk WGS and PTATO accurately detected the known copy number gains and loss (Figures S11C and S11D).

After optimization of SV detection in PTA-based WGS data, we looked for the presence of somatic SVs in the HSPCs of the other patients with FA. We did not observe any large chromosomal abnormalities or translocations (Figure S12). However, we observed 13 deletions with read depth, B-allele frequency (if overlapping germline variants) and split read/discordant read pair support in the 10 cells with sufficient quality (two cells had insufficient quality for accurate CNV detection, Figure S12) ranging from 41 to 29850bp (Figures 5G-5I and Table S3). The deletions were detected in both the PTA-amplified HSPCs as well as the clonally expanded HSPCs, indicating that the detected deletions are probably not artefacts. Additionally, we rarely observed deletions larger than 100bp in the healthy HSPCs sequenced after clonal expansion or PTA, further supporting that there is an increased burden of deletions in HSPCs of FA patients (Figures 5G-5I).

Discussion

The introduction of PTA greatly improved the accuracy of single-cell WGA, leading to rapid adoption in the field[17,18,36–38]. However, bioinformatic tools making optimal use of the potential of PTA have been lacking. To address this, we developed the PTATO pipeline that can accurately distinguish true positive single base substitutions, indels and SVs from false positive artefacts in PTA-based WGS data. The main benefit of PTATO over other tools, in addition to SV filtering, is the relatively high sensitivity between 70%-89% (compared to ~46% reported by SCAN2) to distinguish true base substitutions from artefacts in the callable genome. This means that less extrapolation is required to estimate the true somatic mutation burden in cells, which may be especially important for driver mutation detection and retrospective lineage tracing experiments. The RF model used here was trained and tested mainly on hematological samples, but we showed that it can also effectively remove PTA artefacts from other cell types such as intestinal organoid samples. Nevertheless, if necessary, the RF model included in PTATO can be easily retrained (e.g., by altering the sequence contexts of the true positive variants in the training set as in Figure S3C), making it a flexible tool.

We demonstrated the performance of PTATO by analyzing the genomes of single HSPCs of patients with FA, which could not be clonally expanded *in vitro* for bulk WGS. This analysis showed that most HSPCs of patients with FA have similar somatic mutations burdens as HSPCs of healthy donors, but with an increased number of deletions. These results are in line with findings in mouse models[12] and cell lines[23] of FA. Furthermore, the patterns of SVs detected in the HSPCs of FA patients (mostly deletions <100kb) are similar to the SV patterns found in leukemias[31] and head-and-neck cancers[32] of patients with FA. The increased deletion burden suggests an increased occurrence of double stranded breaks and/or incorrect repair of these breaks in FA HSPCs, which fits with the molecular functions of the FA DNA repair pathway[8]. It is likely that there is selection against HSCs with more genomic rearrangements without the necessary driver mutations to survive, leading to a gradual depletion of such HSCs in FA patients. The analyzed HSPCs of one FA patient with germline *FANCD2*/*BRCA2* mutations showed strongly elevated somatic mutation rates, which is consistent with the broader role of BRCA2 independent of the FA DNA repair pathway[39]. This also highlights that the phenotypic heterogeneity between FA patients may be accompanied by genomic heterogeneity in HSPCs between patients[40]. Further studies including larger patient cohorts are required

to characterize this genomic heterogeneity, which is likely dependent on the causative germline mutations and disease progression stage.

We showed that our PTATO filtering approach improves the usability of PTA, further narrowing the gap in data quality between single-cell WGS and regular bulk WGS. This will be especially important for the genomic analyses of cells that cannot be clonally expanded for regular WGS, such as diseased or differentiated cells. The accurate characterization of single-cell whole genomes by PTA followed by PTATO analysis enables the study of ongoing mutational processes in tissues and cancers, because this combined approach is not limited to analysis of relatively early, clonal mutations like regular bulk WGS[41]. We foresee that such single-cell genome analyses made possible by PTATO will yield an unprecedented view of tumor heterogeneity and cancer evolution.

Limitations of the study

PTATO can detect base substitutions and indels with higher sensitivity (70-89% for callable genomic loci) than other tools like SCAN2 with similar precision (70-92%). The accuracy of somatic variant filtering is generally lower in samples with relatively low mutation burdens (<200 somatic base substitutions) compared to samples with higher burdens, but also for such samples PTATO is more effective in removing PTA artefacts than SCAN2. This illustrates the general challenge to filter mutations in single cells with low mutation burdens such as umbilical cord blood samples, but most cells have more than 200 somatic variants. We note that in some of the analyses performed to determine the performance of PTATO the exact number of PTA artefacts was unknown. In some of these experiments we therefore relied on a mutational signature refit to estimate the number of PTA artefacts in a sample, which is less accurate than using a golden truth set of PTA artefacts. Our strategy to remove indel artefacts based on recurrency and presence in long homopolymers is highly effective in removing PTA indel artefacts, but also excludes some true indels (including some potential disease-causing indels) that are present in bulk WGS samples (Figure 1H). Finally, PTATO enables SV filtering of PTA-based WGS data. Most SV artefacts are removed by PTATO, but the accuracy of SV detection is dependent on the quality of the PTA reaction. Samples with a relatively low DNA output after PTA may show noisy copy number profiles and large regions of loss of heterozygosity due to uneven amplification of the alleles. PTATO calculates quality control metrics to identify such samples with low amplification quality. Precise calculation of performance

metrics (e.g. sensitivity and precision) of SV detection by PTATO will require more WGS data of PTA and bulk samples containing the same SVs.

Acknowledgments

We are grateful to the donors for their participation in this study. We would like to thank Agnes Vissers, Edwin Sonneveld and the biobank of the Princess Máxima Center for their assistance with inclusion of the donors. We thank the Hartwig Medical Foundation (Amsterdam, the Netherlands) for facilitating WGS. This research was supported by grants from the Dutch Cancer Society (KWF Research Project 12682) and European Research Council (ERC; no. 864499) to R.v.B. and The New York Stem Cell Foundation. R.v.B. is a New York Stem Cell Foundation – Robertson Investigator.

Author contributions

S.M., M.E.B. and R.v.B. conceived and designed the study. S.M., F.M. and M.v.R. developed the PTATO computational pipeline. S.M., M.v.R. and F.M. performed computational analyses and designed the figures. F.P., E.J.M.B., I.v.d.W., L.L.M.D., L.T., A.M.B. and S.M. performed sample collection and single cell isolations using flow sorting. D.M.G. performed SCAN2 variant filtering. N.M.G, M.V., L.T. and S.M. generated the AHH-1 cell lines used in this study. C.P. cultured and harvested intestinal organoids. S.M., L.T., N.M.G and M.V. performed PTA. M.E.B., M.B., E.A., D.R. arranged inclusion of the donors and collection of donor material. The manuscript was written by S.M., F.M. and R.v.B. with contributions from all authors.

Declaration of interests

The authors declare no competing interests.

STAR Methods

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.vanBoxtel@prinsesmaximacentrum.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Raw whole genome sequencing data (BAM files) derived from human samples have been deposited at the European Genome-Phenome Archive (EGA) under accession number EGAS00001007288. They are available upon request if access is granted. Details on how to request access are available in the EGA repository. Additionally, de-identified somatic mutation data have been deposited at Mendeley Data (http://dx.doi.org/10.17632/c3r9chw9rb.1) and are publicly available as of the date of publication. Original western blot images have also been deposited at Mendeley Data and are publicly available as of the date of publication. The accession numbers are listed in the key resources table.

All original code has been deposited at Github and is publicly available as of the date of publication. PTATO is freely available as open-source software (https://github.com/ToolsVanBox/PTATO). Code used to analyze the data and create the figures is available at Github (https://github.com/ProjectsVanBox/PTATO).

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

Bone marrow samples were obtained from the biobank of the Princess Máxima Center for Pediatric Oncology with ethical approval under proposal PMCLAB2018-007 and PMCLAB2019-027. Written informed consents from the included individuals were obtained by the Princess Máxima Center. The use of material for this study was approved by the Biobank and Data Access Committee of the Princess Máxima Center. The umbilical cord blood sample of donor CB15 was obtained via the University Medical Center Utrecht (UMCU). The collection of cord blood samples was approved by the Biobank Committee of the UMCU (protocol number 19-737). Informed consent for these samples was obtained by the UMCU. The samples from IBFM26 and IBFM35 were obtained from the German Society of Pediatric Oncology and Hematology (GPOH), who also obtained informed consent from these individuals. Details about the sex and age of the included sample donors can be found in Table S1.

Culture of primary human HSPCs

HSPCs sorted for clonal expansion were cultured in HSPC culture medium for 4 to 7 weeks at 37°C in 5% CO_2 before collection. HSPC culture medium consisted of StemSpan SFEM medium (STEMCELL Technologies) supplemented with SCF (100 ng/ml), FLT3 ligand (100 ng/ml), IL6 (20 ng/ml), IL3 (10 ng/ml), TPO (50 ng/ml), UM729 (500 nmol/l), and Stemregenin (750 nmol/l). Additionally, mesenchymal stromal cells (MSCs) were cultured from a fraction of bone marrow aspirates by plating cells in 12-well culture dishes with DMEM-F12 medium (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum. The medium was refreshed every other day to remove nonadherent cells, and MSCs could be harvested when confluent (after approximately 2 to 3 weeks).

Generation of gene knockouts in AHH-1 cells

Human B-lymphocyte AHH-1 (CRL-8146) cells (male) were purchased from ATCC. Cells were cultured in RPMI 1640 GlutaMAX medium (Thermo Fisher Scientific) supplemented with 1% Penicillin-Streptomycin (Thermo Fisher Scientific) and 10% horse serum (Thermo Fisher Scientific). Guide RNAs (*FANCC*: 5′-GCAAGAGATGGAGAAGTGTA-3′ and *MSH2*: 5′-GTGCCTTTCAACAACCGGTTG-3′) were cloned into pSpCas9(BB)-2A-GFP (PX458) vector (Addgene #48138)[42]. AHH-1 cells were transfected using Lipofectamine 2000 (Thermo Fisher Scientific). One to two days after transfection, GFP-positive transfected cells were single-cell

sorted for clonal expansion on a SH800S Cell Sorter (Sony), which was also used for subsequent clonal steps.

MSH2 inactivation was confirmed using western blot, Sanger sequencing and WGS. The following antibodies were used for western blotting: rabbit anti-MSH2 (D24B5, 1:2000, Cell Signaling Technology) and mouse anti-α-Tubulin (T5168, 1:5000, Sigma-Aldrich). Anti-rabbit IgG IRDye 800CW (1:10000, Li-Cor) and anti-mouse IgG IRDye 680RD (1:10000, Li-Cor) were used as secondary antibodies. Western blots were imaged on an Odyssey DLx imaging system (Li-Cor).

FANCC inactivation was validated by Sanger sequencing, WGS and MMC sensitivity assay. TIDE[43] analysis of the Sanger sequencing traces was performed to estimate indel frequencies in the *FANCC* alleles in the edited cells. For the MMC assay, 5000 cells were plated per well (96-well plates) containing 100µl medium supplemented with different concentrations (0, 5, 10, 50, 100, 500 and 100 nM) of MMC (Sigma-Aldrich) in triplicate. After 5 days of incubation, cell survival was measured using the CellTiter-Glo Luminescent Cell Viability Assay (Promega) according to the manufacturer's protocol.

For the *MSH2*$^{-/-}$ clonal line, two additional consecutive clonal steps were performed (after 48 and 36 days in culture, respectively), and single cells were sorted for PTA 47 days after the third clonal step (Figure 2A). For the *FANCC*$^{-/-}$ clonal line, a second clonal step was performed 58 days after the first clonal step, and PTA was performed 56 days after the second clonal step (Figure 2A). Four clonal lines were generated for the wildtype cells (Figure 2A). From these four clones, two underwent an additional clonal step (43 and 69 days after the first clonal step) and two were single cell sorted for PTA (84 and 87 days after the clonal step). Cells were harvested for DNA extraction when (sub-)clonal lines were sufficiently expanded after single cell sorts.

Intestinal organoid culture

The clonal wild-type human intestinal organoid line ASC-5a from donor STE0072 (female) was derived in a previous study[6]. Intestinal organoids were cultured as previously described[44] in 10 µl domes of Cultrex Pathclear Reduced Growth Factor Basement Membrane Extract (BME) (3533-001, Amsbio) in growth medium consisting of Advanced DMEM/F12 (Gibco), 1× B27, 1× glutamax, 10 mmol/l HEPES, 100 U/ml penicillin-streptomycin (all Thermo Fisher), 1.25 mM N-acetylcysteine, 10 µM nicotinamide, 10 µM p38 inhibitor SB202190 (all Sigma-Aldrich) and the following growth factors: 0.5 nM Wnt surrogate-Fc fusion protein, 2% noggin

conditioned medium (both U-Protein Express), 20% Rspo1 conditioned medium (in-house), 50 ng/ml EGF (Peprotech), 0.5 µM A83-01, and 1 µM PGE2 (both Tocris). For the last two passages, organoids were cultured in medium without antibiotics for 4 days. They were exposed to 0.05% (w/v) FastGreen dye (Sigma) apically, and 5 µg/ml of gentamicin (Sigma) for three days. Primocin (1X, InvivoGen) was added for three days prior to passage or single cell isolation. Single cells were isolated for PTA by dissociating organoids with TrypLE express (Gibco) followed by fluorescence-activated cell sorting (FACS) on an SH800S Cell Sorter (Sony).

METHOD DETAILS

Flow cytometry

Lin^- $CD34^+$ HSPCs were single-cell sorted by fluorescence-activated cell sorting (FACS) on an SH800S Cell Sorter (Sony) for clonal expansion or PTA. The following antibodies were used for staining: CD34-BV421 (clone 561, 1:20), lineage (CD3/CD14/CD19/CD20/CD56)-FITC (clones UCHT1, HCD14, HIB19, 2H7, HCD56, 1:20), CD38-PE (clone HIT2, 1:50), CD90-APC (clone 5E10, 1:200) and CD45RA-PerCP/Cy5.5 (clone HI100, 1:20). AML blasts were selected based on diagnostic immunophenotyping data if available. In most cases, these blasts were CD33, CD38, and/or CD34 positive. All FACS antibodies were obtained from BioLegend.

PTA, DNA isolation and WGS

PTA was performed using the ResolveDNA Whole Genome Amplification Kit (BioSkryb Genomics) according to the manufacturer's protocol. Instead of 10 minutes cell lysis on ice as indicated in the protocol, lysis was performed by 5 minutes incubation on ice followed by 5 minutes incubation at room temperature to maximize DNA denaturation as previously described[36]. DNA samples from bulk AML and bulk MSCs (for germline control) were isolated using the QIAamp DNA Micro Kit (QIAGEN) or DNeasy Blood & Tissue Kit (QIAGEN) according to the manufacturer's instructions. WGS libraries were generated using standard protocols (Illumina). Libraries were sequenced to 15-30x genome coverage (2x150bp) on an Illumina NovaSeq 6000 system at the Hartwig Medical Foundation (Amsterdam, the Netherlands).

WGS read alignment and variant calling

WGS reads were mapped against the human reference genome (GRCh38) using the Burrows-Wheeler Aligner[45] (v0.7.17) mapping tool with settings

'bwa mem –c 100 –M'. Sequence reads were marked for duplicates using Sambamba[46] (v0.6.8). Realignment was performed using the Genome Analysis Toolkit (GATK) (v4.1.3.0)[47]. A description of the complete data analysis pipeline is available at https://github.com/ToolsVanBox/NF-IAP (v1.3.0). Raw variants were called in multi-sample mode by using the GATK HaplotypeCaller and GATK-Queue with default settings and additional option 'EMIT_ALL_CONFIDENT_SITES'. The quality of variant and reference positions was evaluated by using GATK VariantFiltration with options: "--filter-expression 'QD < 2.0' --filter-expression 'MQ < 40.0' --filter-expression 'FS > 60.0' --filter-expression 'HaplotypeScore > 13.0' --filter-expression 'MQRankSum < -12.5' --filter-expression 'ReadPosRankSum < -8.0' --filter-expression 'MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)' --filter-expression 'DP < 5' --filter-expression 'QUAL < 30' --filter-expression 'QUAL >= 30.0 && QUAL < 50.0' --filter-expression 'SOR > 4.0' --filter-name 'SNP_LowQualityDepth' --filter-name 'SNP_MappingQuality' --filter-name 'SNP_StrandBias' --filter-name 'SNP_HaplotypeScoreHigh' --filter-name 'SNP_MQRankSumLow' --filter-name 'SNP_ReadPosRankSumLow' --filter-name 'SNP_HardToValidate' --filter-name 'SNP_LowCoverage' --filter-name 'SNP_VeryLowQual' --filter-name 'SNP_LowQual' --filter-name 'SNP_SOR' -cluster 3 -window 10".

Processing PTA data from external sources

Single-cell PTA-based WGS data (sra files) from cord blood tissue[17] were downloaded from the Sequence Read Archive (accession code SRP178894) and extracted into bam files using the prefetch and sam-dump tools of the sratoolkit (v2.9.2)[48]. Samtools[49] view (v1.3) was then used with the "-bf 1" argument to select for the paired reads and Picard SamToFastq (v2.24.1) was used with the "RG_TAG=ID" and "OUTPUT_PER_RG=true" arguments to generate fastq files. Seqkit[50] replace (v2.2.0) was used to add a sample id to each read name, because they only consisted of a single read number and a number indicating whether it is the first or second read in the pair. Read alignment and variant calling were then performed as described above.

PTATO Nextflow implementation

PTATO was implemented in Nextflow[51] (v21.10.6.5661). Submodules were containerized and automatically downloaded by a container engine, allowing for an easy installation. A Docker image is provided for installation. Singularity (v3.8.7-1.el7) was used for this manuscript, though Docker will also work with a small change to the config. A full PTATO pipeline run (including base substitution filtering, indel filtering and SV calling) required 100-200 CPU hours per sample sequenced to a mean genome coverage of ~15X.

PTATO resources

Next to the sample specific inputs, several general resource files were also used to run PTATO, which are listed in PTATO's "resources.config" file. To make PTATO easy to install and more reproducible, these resource files are included with downloads of PTATO. First, the fasta file and accompanying indexes of the hg38 version of the human reference genome were downloaded from GATK (https://gatk.broadinstitute.org/hc/en-us/articles/360035890811). The input files necessary for the COBALT, GRIDSS2, and GRIPSS tools were downloaded from the Hartwig Medical Foundation (https://nextcloud.hartwigmedicalfoundation.nl/s/LTiKTd8XxBqwaiC?path=%2FHMFTools-Resources)[34,35,52]. A text file containing the centromere locations was downloaded from the UCSC (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=1424951119_QTS0nx5NshNSyspI7KDoJbVh9tci&clade=mammal&org=Human&db=hg38&

hgta_group=map&hgta_track=centromeres&hgta_table=0&hgta_regionType=genome&position=chrX%3A15%2C560%2C138-15%2C602%2C945&hgta_outputType=primaryTable&hgta_outFileName=)[53]. A text file with the genomic coordinates of cytobands was also downloaded from the UCSC (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=1424951119_QTS0nx5NshNSyspI7KDoJbVh9tci&clade=mammal&org=Human&db=hg38&hgta_group=map&hgta_track=cytoBand&hgta_table=0&hgta_regionType=genome&position=chrX%3A15%2C560%2C138-15%2C602%2C945&hgta_outputType=primaryTable&hgta_outFileName=). A bed file with the genomic coordinates of simple repeats was downloaded from the UCSC for hg19 (http://genome.ucsc.edu/cgi-bin/hgTables?db=hg19&hgta_group=rep&hgta_track=simpleRepeat&hgta_table=simpleRepeat). A bed file with the genomic coordinates of gene bodies was downloaded from Ensembl for hg19[54]. A bed file with replication timing data was generated as described previously[6]. Files for which hg19 versions were downloaded were converted to hg38 using UCSCs LiftOver tool[53]. Shapeit maps for hg38 were included with Shapeit (v4.2.2)[55]. Shapeit reference haplotype vcf files were downloaded from the 1000 genomes project (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/).

WGS quality control

Aligment summary metrics were generated for each sample using the CollectAlignmentSummaryMetrics tool from GATK (v4.1.3.0), while WGS metric files were generated using GATKs CollectWGSMetrics tool. Both tools were run using standard parameters. Next, the output of both tools was merged

between all the samples and between the tools using R (v4.1.2). Finally, the R ggplot2[56] (v4.3.0) package was used to generate quality control figures which are combined in a single pdf.

Somatic base substitution and indel filtering

The PTATO pipeline uses a multi-sample VCF file from a single individual and single bam files for each sample (including at least one germline control sample) as input. Preferably, the control samples are analyzed by bulk WGS, as we noted that removal of germline variants can be insufficient when using PTA-based WGS samples as controls. The somatic variant filtering tool SMuRF (https://github.com/ToolsVanBox/SMuRF), which is included in the PTATO pipeline, was used to remove germline and low-quality variants by applying several filters as described previously[6]. Briefly, candidate somatic variants were included if they passed the following filters: no evidence in a paired bulk WGS control sample from the same individual; passed by VariantFiltration with a GATK phred-scaled quality score (QUAL) ≥ 100; base coverage of at least 10 (samples with ~30X genome coverage) or 5 (samples with ~15X genome coverage) in the PTA and paired control sample; a mapping quality (MQ) score of >55; and absence of the variant in a panel of unmatched normal human genomes. Additionally, heterozygous and homozygous base substitutions with a GATK genotype score (GQ) lower than 99 or 10, respectively, were removed. Indels with a GQ score lower than 99 in both PTA or paired control sample were removed. Somatic base substitutions with a variant allele frequency of <0.2 (for samples sequenced at ~15X genome coverage) or <0.3 (for samples sequenced at ~30X coverage) were removed. Somatic indels were required to have a variant allele frequency of at least 0.25. *The R-package VariantAnnotation[57] (v1.42.1) was used to import and export VCF files in R.*

The specific WGS samples that were used as paired bulk WGS control samples to remove germline variants are indicated in Table S1 and Table S2. Briefly, for the AML and FA patients we used bulk MSCs as germline reference. For the AHH-1 cell lines, we used a non-clonal bulk sample of the parental cell line as germline reference. For the cord blood samples, we used a clonally expanded HSPC sample from the same donor to remove germline variants. For patient PMCFANC02, no specific germline control sample was available, because the MSCs did not expand in culture. For this patient, we removed germline variants by selecting only variants that were private for each of the three samples.

Variant calling and filtering by SCAN2 was performed using standard settings (including the signature-based rescue step) as described in the manual (https://github.com/parklab/SCAN2/wiki)[18]. The somatic mutation burden estimated by SCAN2 for each sample was obtained from the .log files. Mutations at chromosome 17 were excluded in comparisons between PTATO and SCAN2, because SCAN2 repeatedly crashed while calling mutations on this chromosome.

Allelic imbalance analysis

Before modelling allelic imbalances, variants on each chromosome were phased separately using SHAPEIT[55] (v4.2.2), with the raw vcf file containing all variants as its input. Additionally, the "sequencing" argument was used, SHAPEIT maps for the relevant reference genome were supplied to the map argument and a vcf with reference haplotypes was supplied to the reference argument.

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

Selection of sequence context features

For each candidate somatic variant, the surrounding 10bp sequence context and mutation type were retrieved using functions modified from the MutationalPatterns R-package[58]. The "closest" function from bedtools[59] (v2.30.0) was used to identify the genes and simple repeat regions closest to the position of each candidate variant. Bedtools merge (with arguments "-d -1 -o min") was used to ensure that each mutation is linked to only one feature of each

feature list. To identify the transcriptional strand bias and replication timing for each somatic mutation, bedtools was used with the "intersect" argument. Some mutations were linked to multiple overlapping gene annotations. For the transcriptional stand bias this was solved by using bedtools with the "merge -d -1 -o distinct" arguments to check if a variant was present in the plus strand, minus strand or both. For the replication timing bedtools was used with the "merge -d -1 -o median" arguments to merge mutations that are present in multiple genes. Next, to merge the gene body, simple repeat, transcriptional strand bias, and replication timing features for each variant, bedtools was used with the "intersect" argument, after which the variants were merged using bedtools with the "merge -d -1 -o unique" arguments.

Linked read analysis using read-backed phasing

For each heterozygous candidate somatic variant, all sequencing reads overlapping the position of the variant were extracted from the sample's bam file. Additionally, all heterozygous germline variants within the area spanned by the reads are extracted from the original input vcf. Next, for each germline variant each read that spans both the germline and somatic variant is checked. Each read that contains either the alternative alleles for both the germline and somatic variant or the reference alleles for both the germline and somatic variant is counted as a cis read. Other reads are counted as trans reads. If a candidate is real, then it would be expected that almost all reads are either cis or trans. Whether the variants are cis, trans, or mixed is then calculated based on a Bayesian likelihood score similar to the one used by SVTyper[60]. The likelihood scores of the three options are then combined into a single Phred-scaled quality score. Candidate variants with a score of <100, between 100-1000 and >1000 were considered to be false positive, uncertain or true variants, respectively.

Random forest training

To obtain a set of true positive variants for training the RF model, base substitutions were selected that were detected in PTA samples of IBFM26, IBFM35, PB10268 and PMCAHH1-FANCCKO and also in bulk WGS-analyzed samples from the same individuals (Figure 1B, Table S1 and Table S2). Somatic base substitutions with a linked read analysis score below 1 in these samples were included in the set of artefacts. Variants that were shared between PTA and bulk WGS samples and also had a linked read analysis score of less than 1 were excluded from both the true positive and the artifact datasets. Variants overlapping with copy number variants and regions of loss-of-heterozygosity in

samples of IBFM26, IBFM35 and PMCAHH1-FANCCKO were excluded from training. Additionally, unique base substitutions detected in three umbilical cord blood HSPCs of donor PMCCB15 analyzed by PTA were considered artifacts, as the number of true mutations in the cord bloods is expected to be very low (20-50)[27]. Finally, the number of base substitutions in the artefact set was subsampled to be the same as the number of base substitutions in true positive set to result in a better class balance.

A random forest was trained on the previously described features with the randomForest (v 4.7-1) R package supplying the "mtry" argument with a value of 4. For some variants (<5%), no p-value for the allelic imbalance or no replication timing value could be calculated (Figure S3D) and therefore they were excluded from the training. To be able to classify variants for which allelic imbalance or replication timing cannot be determined, two additional random forest models were trained: one without the allelic imbalance variable and one without both the allelic imbalance and the replication timing variables. The probability scores calculated by the three RF models were highly correlated (Figure S3E), showing that the additional RF models can effectively classify variants for which allelic imbalance or replication timing could not be determined.

The importance of each variable in the RF model was determined in two complementary ways. First, the mean decreases in Gini coefficient, which is a measure of the contribution of each feature to the homogeneity of the nodes and leaves in the random forest, were obtained from the standard output of model training by the randomForest R-package. Second, to test the impact of each feature on the performance of the model, performance was determined after consecutively removing the feature with the lowest mean decrease in Gini coefficient from the RF model. Each resulting RF model with decreasing numbers of features was applied to the training set to calculate the effect of removing the features one-by-one on the balanced accuracy (true positive rate plus true negative rate divided by 2) of base substitution classification (Figure S1B).

Candidate variant classification by PTATO

For each candidate somatic base substitution, PTATO's main RF model was used to calculate a probability score to predict if a variant is a PTA artefact. A higher score indicates a higher probability that a variant is an artefact according to the RF. For less than 5% of the variants, the allelic imbalance or replication timing could not be determined (Figure S3D). For these variants, the probability scores

of the second (without allelic imbalance) or third (without allelic imbalance and replication timing) RF model were used. Subsequently, two methods were used to determine a sample-specific cutoff value (variants above the cutoff were considered to be artefacts).

First, for each sample a group of likely true positive variants and a group of likely artefacts were selected by taking the variants with either a high (>=1000) or low (<1) linked read analysis score. These variants classified by the linked read analysis were used to validate the performance of the RF model. Precision and recall were calculated for a range of prediction score cutoff values (between 0 and 1 with increments of 0.01). The optimal linked read analysis cutoff was determined by taking the intersection of the precision-recall curves.

Second, a range of different cutoff values (from 0.1 to 0.8 with increments of 0.025) was taken and for each of these cutoffs the variants with a probability score below the cutoff were selected (leading to 29 groups of mutations). For all these 29 groups of mutations, a 96-trinucleotide mutation matrix was calculated using MutationalPatterns[58]. Subsequently, the cosine similarities between all those groups were calculated using the calc_cosim_mutmat() function from MutationalPatterns. Hierarchical clustering of the cosine similarities was performed using the hclust() function in R (Euclidean distance with complete linkage) to generate two clusters: one cluster with low PTA probability cutoffs (and mostly true positives) and one cluster with relatively high cutoffs (and mostly false positives). The highest cutoff value in the cluster with true positives was taken as the cosine similarity cutoff.

Finally, the linked read analyses cutoff and cosine similarity cutoff were merged into a final cutoff that was used to classify variants as true or false positive. This was done by taking the mean of both cutoffs, or by only selecting the cosine similarity cutoff if the highest precision-recall value of the linked read analysis cutoff was below 0.7 (for example in case there were too few variants classified by the linked read analysis).

Somatic indel filtering

Candidate somatic indels were filtered based on recurrency in 139 PTA-based single-cell WGS samples of 22 unrelated individuals. For each included individual, indels occurring in bulk WGS data of the same individual were removed. Subsequently, all remaining somatic indel calls (genomic position, REF and ALT fields from the VCF files) from the PTA-WGS samples with a VAF

>0.15 were collected in a MongoDB database. Indels occurring in at least two different individuals were exported from the database to the PTATO indel exclusion VCF file, which also contains the sample and individual counts and frequencies for each indel. Candidate indels in test samples that overlap with indels present in the exclusion VCF file were removed using the findOverlaps function of the GenomicRanges R-package (v1.48). Additionally, insertions in 5bp+ homopolymers were removed. For this, MutationalPatterns was used to determine the indel type and sequence context around candidate indels.

Mutation burden and signature analysis

The mutational patterns and signature analyses were performed using MutationalPatterns (v3.6.0)[58]. Mutational signatures were used from COSMIC (v3.2) as well as the previously described HSPC, PTA, and ENU signatures[18,26,27,61]. The fit_to_signatures_bootstrapped function of MutationalPatterns (with parameters n_boots=100 and method="strict") was used to perform strict mutational signature refitting. Figures were made using ggplot2 (v3.4.1)[56].

CallableLoci from GATK v3.8.1 (with parameters --minBaseQuality 10 --minMappingQuality 10 --minDepth 8 --minDepthForLowMAPQ 10 --maxDepth 100) was used to determine the fraction of the sequenced genome that had sufficient coverage and quality for variant calling. Variants not overlapping with the callable regions determined by CallableLoci were excluded. Subsequently, all remaining variants on autosomal chromosomes were counted. To obtain the mutation burden, the mutation count was extrapolated by dividing it by the fraction of the genome that was surveyed (determined by CallableLoci), as previously described[6].

A linear mixed-effects model was used to correlate the mutation burden in HSPCs from healthy donors and the age of the donors as previously described[28]. This model was used to calculate the expected mutation burdens for the specific ages of the patients. The 95% confidence and 95% prediction intervals were calculated using the R package ggeffects (v1.1.0)[62].

In silico mixing of true and false variants

To determine how well PTATO can classify artefacts in datasets with different numbers of true base substitutions, PTATO was first applied to PTA samples PB15778-DX1BM-HSCPTAP1D12, PB32346-DX1BM-HSCPTAP3A7 and PMCCB15-CBCMP-PTAP3D10 to calculate the features of each base substitution. To obtain true positive variants, 800 base substitutions that were shared between the

PTA samples (PB15778-DX1BM-HSCPTAP1D12 and PB32346-DX1BM-HSCPTAP3A7) and their corresponding bulk WGS samples (PB15778AMLBULK and PB32346-DX1BM-AMLBULK, respectively) were selected. From these 800 true positive variants, different numbers of variants (ranging from 100 to 800 with steps of 100) were randomly selected and merged with 465 base substitutions of PMCCB15-CBCMP-PTAP3D10 to create datasets with different ratios of true and false positives (with the values of the features from the samples in which the variants were originally detected). PTATO's RF model was applied to each of these datasets to calculate how many true positive variants (variants that originated from samples of PB15778-DX1BM-HSCPTAP1D12 and PB32346-DX1BM-HSCPTAP3A7) remained after filtering and how many artefacts variants that originated from PMCCB15-CBCMP-PTAP3D10) were removed.

To test how well PTATO can classify variants with different mutational backgrounds, the 800 base substitutions shared between the PTA and bulk WGS samples from donors PB15778 and PB32346 were selected. For each mutational signature in the Cosmic Mutational Signatures database v3.2 (https://cancer.sanger.ac.uk/signatures/sbs/), the mutation type and the base up- and downstream features were modified in the feature tables of these 800 selected true base substitutions (while keeping all the other features the same). Subsequently, each set of the 800 true base substitutions with modified mutation spectra was merged with 465 base substitutions from PMCCB15-CBCMP-PTAP3D10. PTATO's RF model was applied to each of these datasets to determine how many true positive variants with modified mutation spectra (variants that originated from samples of PB15778-DX1BM-HSCPTAP1D12 and PB32346-DX1BM-HSCPTAP3A7) remained after filtering and how many artefacts variants that originated from PMCCB15-CBCMP-PTAP3D10) were removed.

Normalization of copy number ratios for SV detection

GC-normalized read depth per 1000 basepair genomic window was calculated by COBALT (v1.11)[35] (Figures S10C and S10D). Cosine similarities between raw genome-wide copy number profiles (1kb resolution) were calculated by using the cos_sim_matrix function of MutationalPatterns. A coverage panel-of-normals (PON) was generated by merging COBALT ratio files of 12 copy number neutral PTA samples. The total read counts from all windows of each sample were first normalized so that every sample has the same total amount of read counts. Subsequently, the mean readcount per bin over all normal samples in the PON was calculated. PTATO uses the coverage PON file to smoothen PTA-specific coverage fluctuations. First, the total read depth in a test sample is

normalized to the same total amount of read counts in the coverage PON. Subsequently, the read counts in each window are divided by the mean read counts in the same window in the PON (Figure S10E). Additionally, the bottom and top 1% outlier windows in the PON file and the windows located within 1Mb distance of centromeres and telomers are excluded from the analysis.

The smoothened read counts were subsequently binned in 100kb windows (Figure S10F). The copynumber (v1.34.0) R-package with parameter "gamma=100" was used to segment the median read count data in both the 100kb and 1kb windows[63] (Figure S10G). The segments based on the 100kb resolution were used as raw copy number segments. The start and end coordinates of these raw copy number segments were fine mapped by taking the start and end coordinates of overlapping 1kb window-based segments.

To determine if the read count distribution within a segment was different from normal diploid segments in a sample, the read counts per 1kb from the top 25% of the segments with a mean copy number closest to 2 in the sample were selected (Figure S10H). For each segment, a Z-score was determined by first subtracting the mean copy number in the segment by the mean copy number in the 25% segments with a copy number closest to 2, followed by dividing this number by the standard deviation of the copy number in the normal segments. The pnorm function in R was used to determine the significance in difference in coverage distributions between the segment and the 25% segments with a copy number closest to 2, which was called the "sample p-value". One-sided tests were used to determine if the copy number in the segment is either higher or lower than the diploid segments.

Each segment was overlapped with the mean read counts per 1kb bin in the coverage PON to compare the coverage distribution between the sample and the PON in the segmented region (Figure S10H). For each segment, a Z-score was determined by first subtracting the mean copy number in the segment by the mean copy number in the PON, followed by dividing this number by the standard deviation of the copy number in the segment in the PON. The pnorm function in R was used to determine the significance in difference in coverage distributions in the segment between the test sample and the PON, which was called the "PON p-value". One-sided tests were used to determine if the copy number in the segment is either higher or lower than in the PON.

The segments with a sample p-value <0.05 and a PON p-value <0.2 were considered as potential copy number gains or losses in the later filtering steps that integrate the coverage and B-allele frequency segments.

Deviation of allele frequency calculations

VAFs of germline variants can be noisy in PTA-based WGS data due to uneven genome amplification, which impedes accurate copy number variant detection based on raw B-allele frequencies. To reduce noise due to uneven amplification, the VAFs of germline base substitutions were first binned in 100kb windows instead of taking separate B-allele frequencies of each individual variant. To determine a mean allele frequency for multiple variants in a bin, the deviation of allele frequency (DAF) was calculated by taking the absolute value after subtracting the VAF of each variant from 0.5 (which is the expected VAF for a perfectly amplified and sequenced germline variant). Thus, each variant has a DAF between 0 (corresponding to a VAF of 0.5) and 0.5 (corresponding to a VAF of 0 or 1). Subsequently, all DAF values of germline base substitutions are binned in 100kb genomic regions and the mean DAF for each region is calculated (Figure S11A). The copynumber R-package with parameter "gamma=100" was used to segment the 100kb bins in crude DAF regions (Figure S11A). These crude segments were fine mapped by adjusting the start and end coordinates of the segments to the positions of the nearest germline SNVs (within 200kb distance of the segment) with similar DAFs as the segment.

Binning and segmenting were performed partly different to detect segments of potential copy number gains (Figure S11A). A small portion of genomic loci displayed loss-of-heterozygosity (LOH) because one of the alleles was not properly amplified by PTA. These artificial LOH regions may especially affect detection of copy number gains, because these regions have a relatively high DAF. Therefore, in parallel to binning and segmenting DAFs for detection of cnLOH and copy number losses as described above, PTATO also performed binning and segmenting after exclusion of all germline variants with a DAF >0.45 (corresponding to LOH) for detection of copy number gains (Figures S11A and S11B). Thus, PTATO determined two types of segments: one group of segments based on all germline variants for detection of copy number losses and cnLOH regions, and one group of segments based on only germline variants not displaying LOH for detection of copy number gains.

Finally, also the distribution of the VAFs of each germline variant was taken into account for CNV detection. The VAFs of germline variants in a normal diploid

segment have a unimodal normal distribution around VAF=0.5 (Figure S11B). In contrast, VAFs of germline variants in segments with copy number losses or gains are expected to have a bimodal distribution with modes at 0 and 1 for copy number losses and modes at 0.33 and 0.66 for copy number gains (with a copy number of 3) (Figure S11B). Therefore, PTATO used the Modes() function from the LaplacesDemon R-package (v16.1.6) to calculate the modes of the VAF distributions in each segment.

The segments with a DAF of more than 0.45 (corresponding to VAF < 0.1 or > 0.9) were considered to be LOH regions in the following integration of copy number segments and DAF segments (Figure S11B). The segments calculated after exclusion of LOH variants were used to select potential copy number gains. From these segments, only the segments that had 1) a mean DAF more than the average DAF in the sample and 2) more than one VAF distribution mode, of which one should be around 0.33 (+/- 0.12) and one should be around 0.66 (+/- 0.12), were selected as potential copy number gains (Figure S11B).

SV breakend calling and filtering

Somatic SV breakends were called by GRIDSS v2.13.2 and prefiltered by GRIPSS v1.9 using a corresponding bulk-sequenced germline control[34]. StructuralVariantAnnotation v1.12.0 was used to import and export SV vcf files in R. The GRIPSS-filtered somatic breakends of 15 PTA-based samples of four unrelated individuals were merged using bedtools[59] merge (v2.30.0). Breakend positions occurring within 2000bp of each other in multiple of these individuals were included in a breakend PON. Candidate breakends in other samples overlapping with the regions in the breakend PON were removed. Subsequently the normalized coverage and DAF of the SV candidates was calculated. Breakends of duplications were filtered if the DAF was less than 0.18 and/or the copy number ratio was <2.5. Breakends of deletions were filtered if the DAF was less than 0.4 and/or the copy number ratio was >1.5. Breakends with a coverage of more than 100 were also excluded for samples with a targeted genome coverage of 15x as many artefacts occur in these regions with excess coverage. Inversions were filtered if they only have one breakpoint junction instead of two. Additionally, all inversions less than 1kb in size were removed. Inter-chromosomal events were also filtered if they only have one breakpoint junction (instead of two), unless they were situated less than 100kb from a copy number variant. This exception rescues unbalanced translocations.

The GRIDSS-PURPLE-LINX pipeline (v1.3.2) developed by the Hartwig Medical Foundation[35] was used for SV calling and filtering in bulk WGS samples.

Integration of coverage, allele frequencies and structural variant breakends

The coverage segments, DAF segments, and breakends of SV candidates were intersected to create the final list of filtered structural variants. Copy number variants were required to have both coverage and DAF support (based on the thresholds described above), but not necessarily breakend support, as many CNVs have start and/or end positions within repeat regions that are difficult to capture with PTA and/or short-read sequencing. Segments with a mean DAF of >0.45 (corresponding to VAFs of <0.1 and >0.9) that did not overlap with coverage segments of copy number losses or gains were considered to be copy number neutral loss-of-heterozygosity (cnLOH) regions. ggplot2[56] and Circos[64] (v0.69-9) were used for to visualize structural variants and karyograms. The SVs that were left after filtering were manually inspected by visualizing the reads in the bam files using the Integrative Genomics Viewer (IGV)[65] for further validation.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests were performed with R and the rstatix and ggpubr R-packages. Details of each test are described in figure legends.

Supplemental information titles and legends

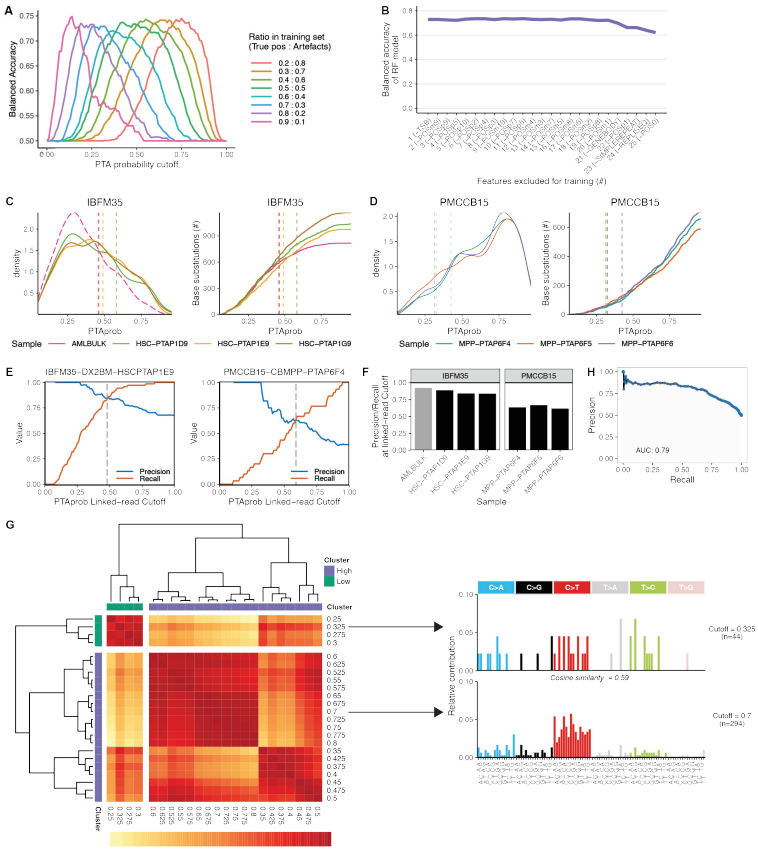


Figure S1. Calculations of optimal PTA probability cutoffs by PTATO, Related to Figure 1. **(A)** Training of the RF model was tested on different ratios of true and false positives to determine the optimal mix of variants. Changing these ratios mainly leads to a shift in PTA probability scores, while the optimal balanced accuracy remains the same (but achieved at a different cutoff value). The model trained at a 1:1 ratio shows the broadest range of cutoffs scores at which an optimal balanced accuracy is achieved, showing that this model is the most robust. **(B)** The effect on accuracy of cumulatively removing features one-by-one for training of the RF model. **(C)** Distributions (left) and cumulative distributions (right) of the PTA probability scores (PTAprob) of candidate base substitutions before PTATO filtering in one bulk WGS (with relatively low PTAprob scores) and three PTA-based WGS samples. Vertical lines indicate the sample-specific PTAprob cutoffs determined by PTATO. **(D)** Same as (C), but then for umbilical cord blood samples with low mutation burdens. **(E)** Precision and recall at different PTAprob cutoffs of a subset of base substitutions that could be classified as true or false positive by the linked read analysis. The linked read cutoff is determined by taking the PTA probability at minimal difference between the precision and recall. **(F)** Overview of the linked read precision-recall rates of samples in the training set. Samples with low mutations burdens

can have low precision-recall rates, as shown here for cord blood donor PMCCB15, which requires an alternative method to calculate an optimal cutoff. **(G)** Heatmap showing the cosine similarities between 96-trinucleotide mutational profiles calculated for different PTAprob cutoffs in sample PMCCB15-CBMPP-PTAP6F4. Hierarchical clustering is used to make one cluster with low PTAprob cutoffs (containing most true positives) and one cluster with high PTAprob cutoffs (containing most artefacts). The highest value in the cluster with true positives is used as the cosine similarity cutoff (0.325 in this case). Two example profiles of the mutation sets at different cutoffs are shown on the right. **(H)** Precision-recall curve showing the performance of the random forest using all input variables on the out-of-bag training data for different probability cutoffs.

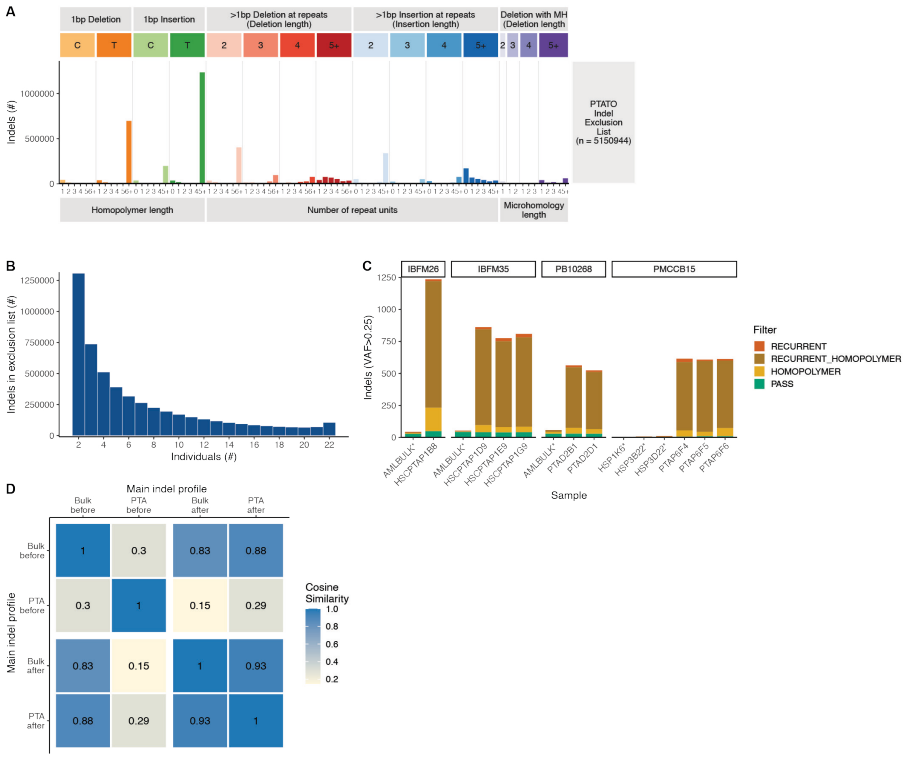


Figure S2. Indel filtering by PTATO based on recurrency and sequence context, Related to Figure 1.

(A) Profile of the indels present in the list of recurrent indels that is used by PTATO to filter indel artefacts. The exclusion list contains mostly insertions at long homopolymers, but also recurrent deletions at long homopolymers. This indicates that just excluding insertions at long homopolymers is not sufficient to remove all indel artefacts. **(B)** Histogram showing in how many individuals (out of 22) the indels in the exclusion list are found. **(C)** PTATO filters indel artefacts by filtering insertions at long homopolymers (HOMOPOLYMER) and by filtering indels recurrent in multiple unrelated individuals (RECURRENT). This filtering removes most excess indels (the remaining indels are labelled with PASS), but also limits sensitivity to detect insertions in long homopolymer tracts. Samples indicated with an asterix (*) are bulk (non-PTA) WGS samples. **(D)** Heatmap showing the cosine similarities between the main indel spectra of the PTA- and bulk-WGS samples (from Figure 1H) before and after filtering by PTATO. The main indel spectra (16 channels) of the 6 PTA- and 3 bulk WGS samples of the AML patients shown in Figure 1H were merged by type (PTA, bulk, before and after filtering) before calculating the cosine similarities.

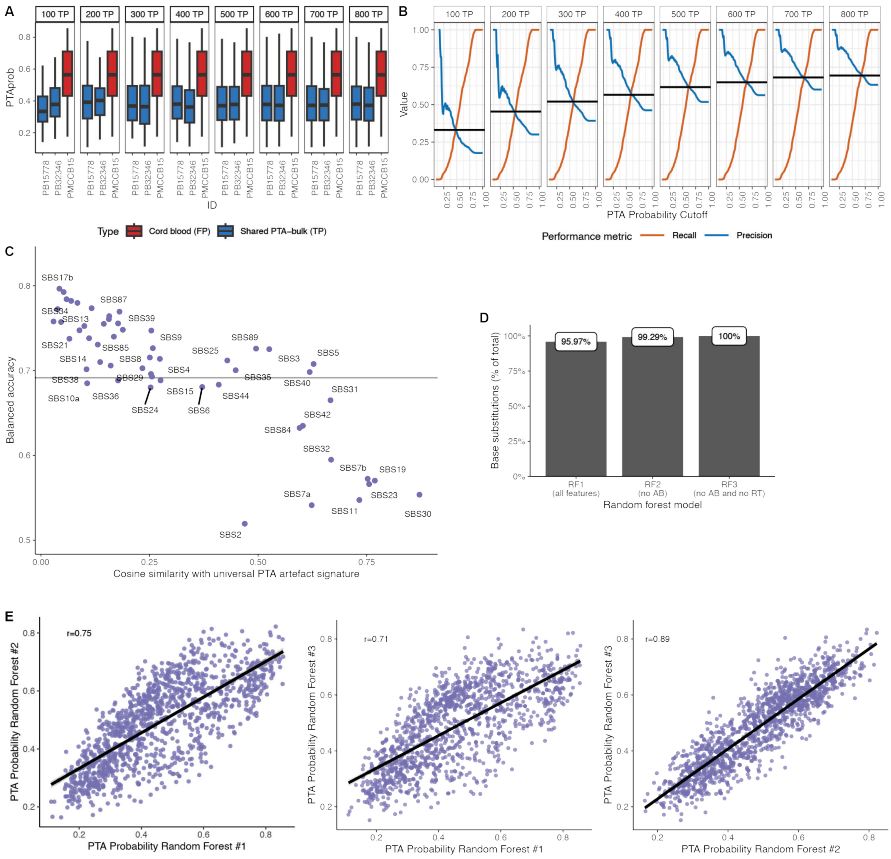


Figure S3. Validating PTATO performance by in silico mixing and mutating base substitutions, Related to STAR Methods.

(A) PTA probability scores calculated by PTATO for base substitutions found in PTA-based WGS samples of two AML patients (PB15778 and PB32346) and one umbilical cord blood sample. For the two AML patients, only likely true positive (TP) base substitutions were included that were shared by the PTA sample and the bulk WGS sample. Most of these shared base substitutions have lower PTA probability scores compared to the base substitutions detected in the cord blood sample (most of which are PTA artefacts). **(B)** Precision-recall curves showing the performance of identification base substitution classification when mixing different amounts of true and false positives. True base substitutions were obtained by selecting mutations shared between PTA and bulk WGS samples of two AML patients. Different numbers of true positives (shown in the headers) were mixed with 465 base substitutions of a cord blood sample, which are considered artefacts. We note that roughly 10% of the base substitutions in the cord blood samples (~50 out of 465) are estimated to be real base substitutions, leading to an

underestimation of the performance. **(C)** Balanced accuracy of PTATO in distinguishing in silico mutated true positive base substitutions from PTA artefacts (465 base substitutions from a cord blood sample). The trinucleotide contexts of sets of 800 base substitutions shared between bulk and PTA-based WGS samples from two AML patients were in silico mutated (while keeping the other RF features the same) to match the profiles of the depicted COSMIC mutational signatures. **(D)** For the majority of the base substitutions analyzed here (n=1265, 800 from the AML samples and 465 from the cord blood sample), all RF features could be determined. For some variants, values for the allelic imbalance (AB) and/or replication timing (RT) variables could not be calculated (for example due to low amplification quality or sequencing depth of the locus). For this small subset of variants, the PTA probabilities of the second or third random forest model (which exclude allelic imbalance and allelic imbalance plus replication timing, respectively) are used to determine if a variant is a PTA artefact. **(E)** Correlations (Pearson) between the PTA probability scores calculated by the first (all features), second (without the allelic imbalance feature) and third (without the allelic imbalance and replication timing features) random forest models for the base substitutions (dots) analyzed here. As only a small number of variants cannot be analyzed by random forest 1 and because the probabilities calculated by the three models are highly correlated, random forest 2 and 3 are only expected to have a minor effect on variant filtering. Nevertheless, they can be useful to rescue the small subset of variants that cannot be analyzed with the primary model.

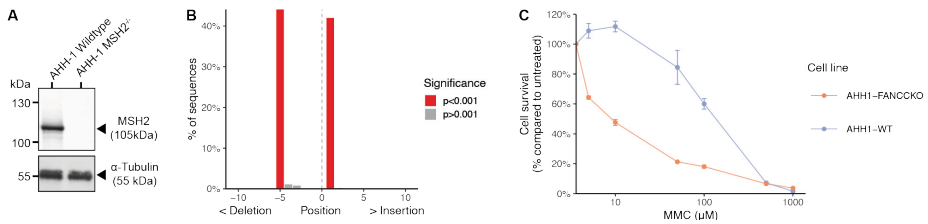


Figure S4. Validation of MSH2 and FANCC knockout status in AHH-1 cell lines, Related to Figure 2.

(A) Western blot showing the absence of MSH2 protein expression in the AHH-1 MSH2-/- clonal cell line. **(B)** TIDE analysis detects a 5-basepair deletion and 1-basepair insertion introduced by CRISPR/Cas9 in the FANCC gene of the AHH-1 FANCC-/- clonal cell line. Due to the absence of high quality antibodies, western blotting could not be performed to study FANCC protein expression. Therefore, we used PCR and Sanger sequencing followed by TIDE decomposition, in addition to a Mitomycin C (MMC) sensitivity assay, to confirm knockout status. The presence of the biallelic indels in FANCC was also confirmed in the WGS data (data not shown). **(C)** MMC sensitivity assay showing the hypersensitivity of the AHH1 FANCC-/- clonal cell line to the DNA cross-linking agent MMC. This finding provides additional support for the knockout status of FANCC in this cell line, as cells of patients with FA are known to display MMC hypersensitivity. Mean survival values from triplicate experiments are shown and error bars indicate standard deviations.

Figure S5. Single base substitution filtering in PTA-based WGS data of AHH-1 cell lines by PTATO, Related to Figure 2.

(A) Number of base substitutions aqcuired per day in culture between single cell steps in subclones analyzed by bulk WGS (green) and PTA samples before (lightblue) and after (darkblue) PTATO filtering. **(B)** Number of unique base substitutions not present in the (sub)clones reported by PTATO and SCAN2 before and after extrapolation. PTATO detects more base substitutions, requiring less extrapolation to estimate the true base substitutions burden in a cell. The horizontal black lines indicate the expected number of base substitutions based on the days in culture since the previous single-cell step and the mutation rate in the corresponding subclones. **(C-E)** The 96-trinucleotide

mutational profiles of the wildtype (WT) (C), FANCC-KO (D) and MSH2-KO (E) AHH-1 cells assessed by WGS after clonal expansion or after PTA. The variant calls before PTATO filtering (RAW) still contain numerous PTA artefacts. The profiles of the variants removed by PTATO are shown in the middle panels (PTA_FAIL). **(F)** 96-trinucleotide profiles of the base substitution signatures extracted by non-negative matrix factorization (NMF). One signature resembles the PTA artefact signature (red), one resembles the background signature for AHH-1 cells (green) and one resembles signatures found in mismatch repair deficient cells (blue). **(G)** Contribution of the signatures extracted by NMF (F) to the mutational profiles of each sample. The mutations removed by PTATO (PTA_FAIL) are mostly refitted to the PTA artefact signature. The mutational profiles of the PTA samples filtered by PTATO and SCAN2 are more similar to the profiles of the subclones analyzed by bulk WGS than the unfiltered (PTA_RAW) samples.

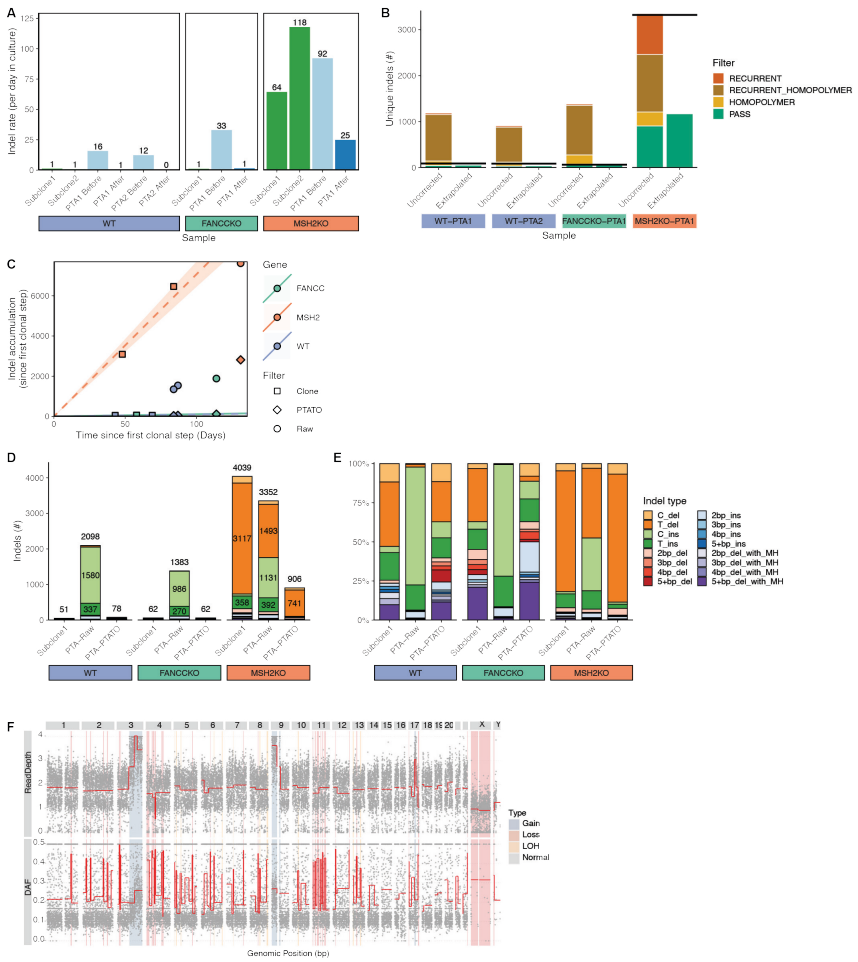


Figure S6. Indel filtering in PTA-based WGS data of AHH-1 cell lines by PTATO, Related to Figure 2.

(A) Number of indels acquired per day in culture between single cell steps in subclones analyzed by bulk WGS (green) and PTA samples before (lightblue) and after (darkblue) PTATO filtering. **(B)** Number of unique indels not present in the (sub)clones reported by PTATO before and after extrapolation. The horizontal black lines indicate the expected number of indels based on the days in culture since the previous single-cell step and the indel accumulation rate in the corresponding subclones. **(C)** Accumulation of indels since the first clonal step. The circles and diamonds indicate the number of indels detected in the PTA samples before and after PTATO filtering, respectively. **(D)** Number of indels (not present in the preceding clonal step) in the subclones analyzed by bulk WGS and the PTA samples before (Raw) and after PTATO filtering. More than a thousand artificial indels are detected in the wildtype and FANCC-/- PTA samples. **(E)** Relative contributions of the different types of indels (not present in the preceding clonal step) detected in the subclones analyzed by bulk WGS and the PTA samples before (Raw)

and after PTATO filtering. PTATO mostly removes 1-basepair (bp) insertions. **(F)** Copy number and deviation-of-allele frequency (DAF) plots of sample PMCAHH1-MSH2KO-C27E06SC51B06-PTAP1E7. This sample has many loss-of-heterozygosity (LOH) regions, indicating a lower quality genome amplification by PTA. MH, microhomology.

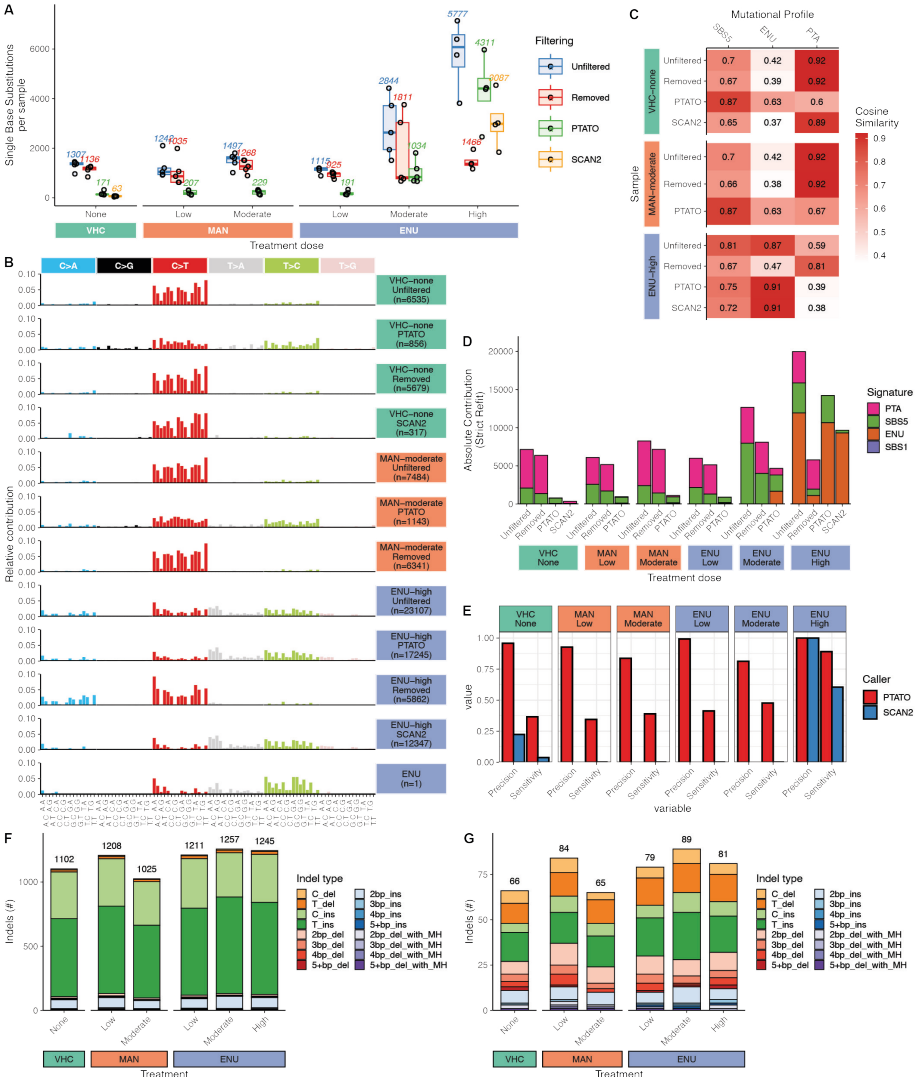


Figure S7. PTATO accurately filters PTA artefacts from a PTA-based WGS dataset of cord blood cells, Related to Figure 2.

(A) Boxplot showing the number of base substitutions for human cord blood samples treated with different concentrations of a vehicle control (VHC; n = 5), D-mannitol (MAN; low: n = 5, moderate: n = 5) or N-ethyl-N-nitrosourea (ENU; low: n = 5, moderate: n = 5, high: n = 4). Numbers above the boxes indicate the mean base substitution burden per sample in each treatment group. **(B)** The 96-trinucleotide profiles of the base substitutions in the indicated treatment groups before ("Unfiltered") or after ("Filtered") PTATO filtering or the base substitutions removed by PTATO ("Removed") or the mutations detected by SCAN2. The bottom panel shows the profile of the mutational signature that has been previously associated with ENU-treatment. **(C)** Cosine similarities of the mutational profiles of the

base substitutions that are present before ("Unfiltered") or after PTATO filtering ("PTATO"), or that are removed by PTATO or detected by SCAN2, with the SBS5-, ENU- and PTA mutational signatures. Variants detected by SCAN2 in the VHC samples show a strong similarity to the PTA artefact signature, suggesting it mostly detects artefacts in these samples. **(D)** Contributions of the PTA, SBS1, SBS5 and ENU mutational signatures to the profiles of the unfiltered, removed and filtered base substitutions determined by a bootstrapped strict mutational refit. PTATO mostly removes mutations associated with the PTA mutational signature, while keeping the mutations associated with SBS5 and the ENU mutational signatures. The base substitutions were pooled for each treatment dose. **(E)** Precision and sensitivity of base substitution detection by PTATO and SCAN2. Precision is defined as 1 minus the fraction of base substitutions refitted to the PTA artefact signature. Sensitivity is defined as mean contribution of SBS1, SBS5 and ENU-signatures in the Unfiltered call sets minus mean contribution of SBS1, SBS5 and ENU-signatures in the PTATO and SCAN2 call sets. **(F)** Mean numbers and types of indels found per sample in each treatment group before filtering by PTATO. **(G)** Mean numbers and types of indels found per sample in each treatment group after filtering by PTATO. PTATO removes over a thousand indels per sample, mainly C- and T-insertions at homopolymers. As has been shown before, treatment with ENU did not cause an increase in indel burden. MH, Microhomology.

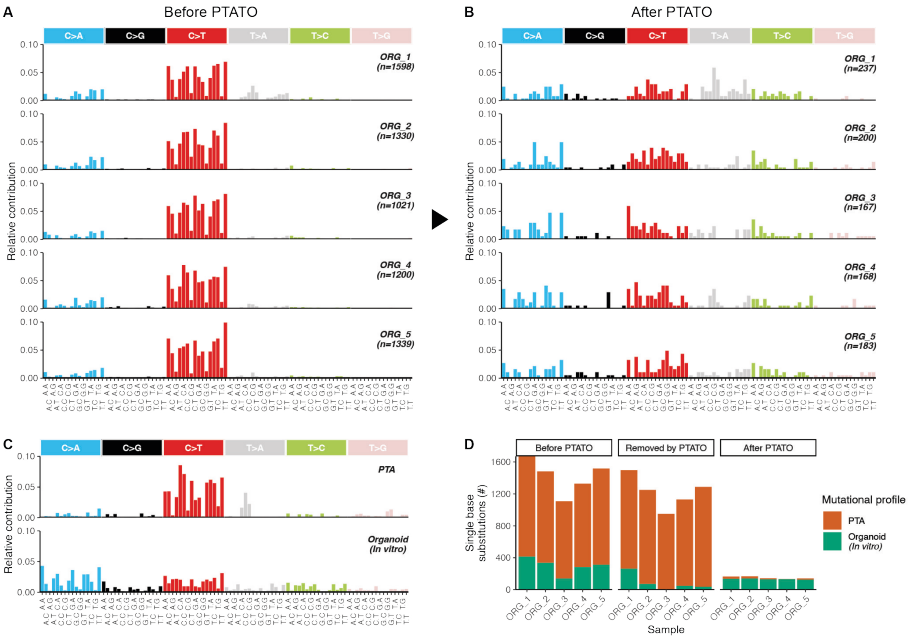


Figure S8. Filtering of PTA-based WGS data of single intestinal organoid cells by PTATO, Related to Figure 2.

(A) The 96-trinucleotide mutational profiles of five single intestinal organoid cells analyzed by PTA-based WGS, before PTATO filtering. **(B)** The 96-trinucleotide mutational profiles of five single intestinal organoid cells analyzed by PTA-based WGS, after PTATO filtering. **(C)** The 96-trinucleotide mutational profiles of the PTA artefact (top) and organoid (bottom) mutational signatures used for signature refitting. The profile of base substitutions that accumulate during in vitro culture of intestinal organoids was previously determined by analysis of the subclonal mutations in WGS data of clonal organoids. **(D)** Contribution of mutational signatures to the base substitution profiles of the five organoid cells determined by bootstrapped signature refitting. Filtering by PTATO removes nearly all base substitutions that could be attributed to the PTA artefact signature, showing that it is also applicable to non-hematological PTA samples.

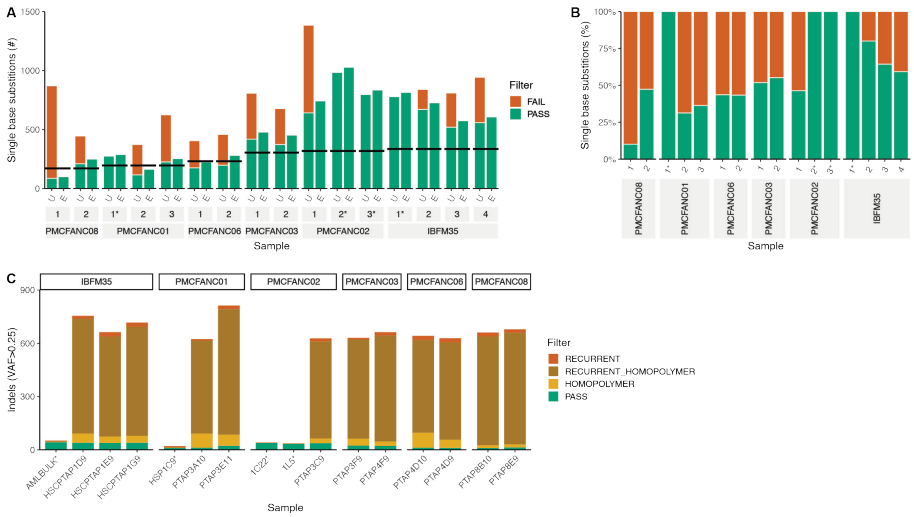


Figure S9. PTATO filtering of single base substitutions and indels in HSPCs of patients with FA, Related to Figure 3.

(A) Absolute number of single base substitutions that passed (PASS) or failed (FAIL) filtering by PTATO, before (U = unfiltered) and after (E = extrapolated) extrapolation based on CallableLoci. The horizontal black lines indicate the expected number of base substitutions for each individual based on their age. Samples not amplified by PTA are marked with an asterisk. **(B)** Relative amount of single base substitutions that that passed or failed filtering by PTATO. **(C)** Number of indels per sample that passed filtering by PTATO or that were removed by PTATO because they are present in the recurrent indel filter list (RECURRENT) or are insertions in homopolymer regions (HOMOPOLYMER), or both (RECURRENT_HOMOPOLYMER).

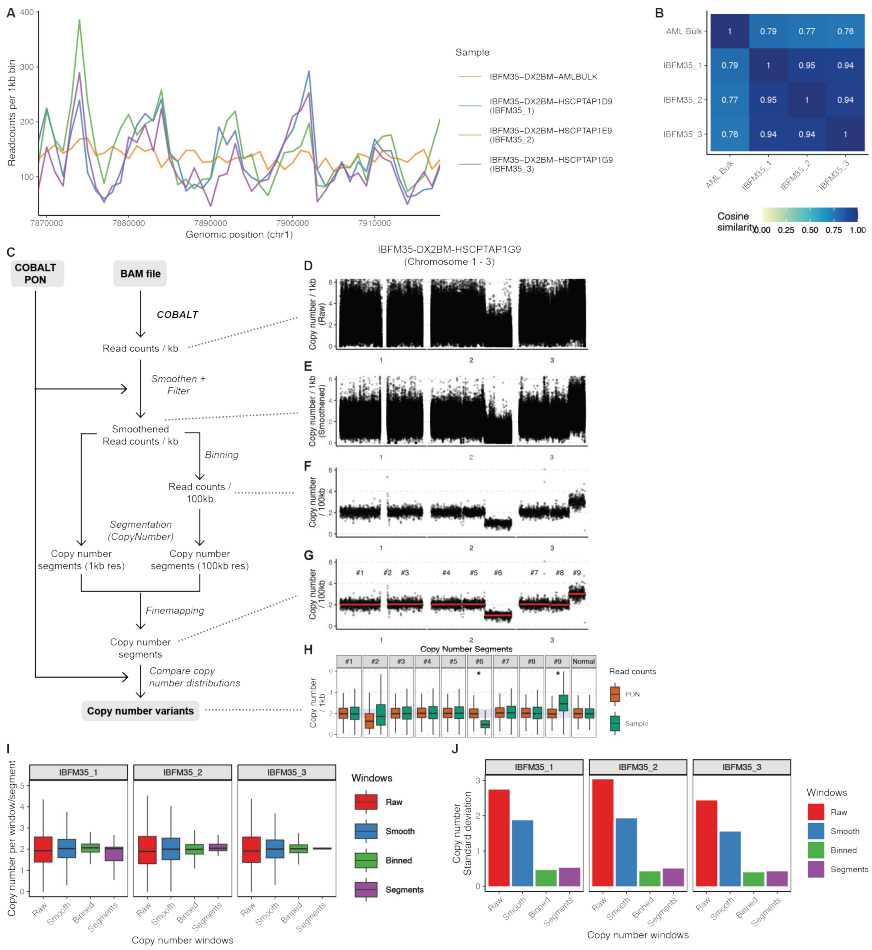


Figure S10. Copy number variant detection by PTATO based on read depth, Related to Figure 5.

(A) Coverage profiles determined by COBALT (at 1kb resolution) of three PTA samples and one bulk WGS samples in a 50kb region on chromosome 1. **(B)** Heatmap showing the cosine similarities between the genome-wide coverage profiles (1kb resolution). **(C)** Overview of the first part of copy number filtering (based on coverage) performed by PTATO. (D) Example of a copy number profile (1kb resolution) of three chromosomes determined by COBALT, before any filtering by PTATO. **(E)** Copy number profile (1kb resolution) after smoothening by PTATO using the PON. **(F)** Copy number profile at 100kb resolution after binning the smoothened read counts by PTATO. **(G)** Copy number profile (100kb resolution) which shows the calculated copy number segments as red horizontal lines. The detected segments are labelled by the numbers above the plot (#1 to #9). **(H)** Distributions of the copy numbers (1kb resolution) in the 12 samples in the PON (containing normal diploid samples) and the test sample (IBFM35-DX2BM-HSCPTAP1G9) for each of the 9 detected segments (on the three chromosomes) with similar copy numbers. Additionally, in the last panel the coverage distributions in the top 25% of the bins closest

to copy number 2 are shown to depict the variation in copy number in regions that are considered to be normal diploid. These coverage distributions were used by PTATO to determine which segments are potentially copy number gains or losses, as indicated by the asterisk. In later steps, these segments of copy number variant candidates were intersected with segments with divergent germline variant allele frequencies to generate the final copy number variant call set. **(I)** The effects of each consecutive coverage filtering step on the variance in copy number between genomic windows. **(J)** The standard deviation of the copy numbers in each genomic window after each coverage filtering step. This shows that each filtering step further reduces the variance in copy number profiles.

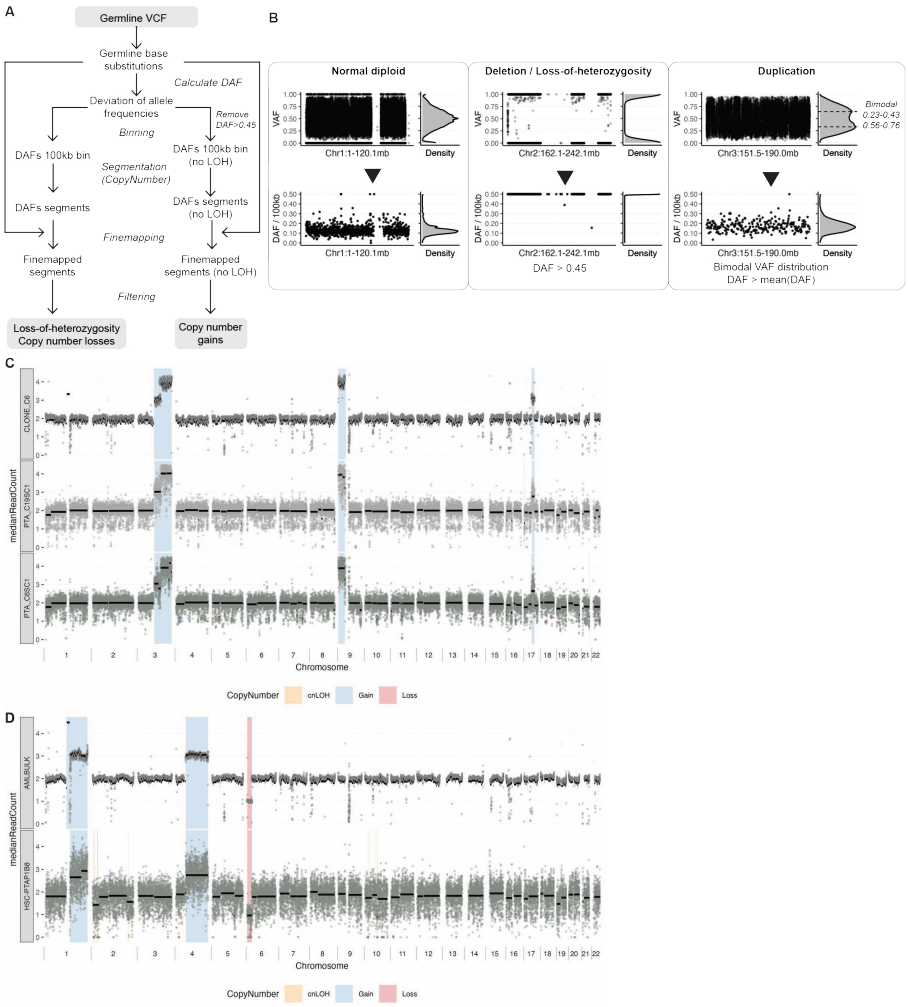


Figure S11. Copy number calling with allele frequencies germline base substitutions, Related to Figure 5.

(A) Schematic overview of the filtering steps performed by PTATO to identify copy number changes based on allele frequencies of germline base substitutions. Filtering for loss-of-heterozygosity (LOH) and deletions on the one hand, and copy number gains on the other hand were performed in parallel. For detection of copy number gains, first all germline variants with a DAF of >0.45 (corresponding to a loss of heterozygosity) were removed. This was done to minimize the effects of LOH that was caused to uneven DNA amplification by PTA on detection of duplicated regions. **(B)** Examples of VAF and DAF distributions of a copy number neutral region (left), a genomic region with a copy number loss (center) and a genomic region with a copy number gain (right) in sample IBFM35-DX2BM-HSCPTAP1G9. These examples depict how PTATO made use of germline variant allele frequencies as a part to identify copy number variants. LOH and deletions

events are called if the mean deviation of allele frequencies (DAF) in a segment was more than 0.45. Duplications were called if the mean DAF of a segment is higher than the mean DAF in the entire sample and if there was a bimodal distribution of the VAFs of germline variants in a segment with modes of ~0.33 and ~0.66. **(C)** Copy number profiles (at 100kb resolution) of one bulk WGS (CLONE_C6) and two single-cell PTA-based WGS samples of clonal AHH-1 cell lines. Colored background shadings show the copy number calls made by PTATO (for the PTA-based WGS samples) or PURPLE (for the bulk WGS sample). The black horizontal lines depict the copy number segments determined by PTATO (for the PTA-based WGS samples) or COBALT (for the bulk WGS sample). **(D)** Copy number profiles (at 100kb resolution) of one bulk WGS (AMLBULK) and one single-cell PTA-based WGS sample (HSC-PTAP1B8) of AML patient IBFM26. Colored background shadings show the copy number calls made by PTATO (for the PTA-based WGS sample) or PURPLE (for the bulk WGS sample). The black horizontal lines depict the copy number segments determined by PTATO (for the PTA-based WGS sample) or COBALT (for the bulk WGS sample).

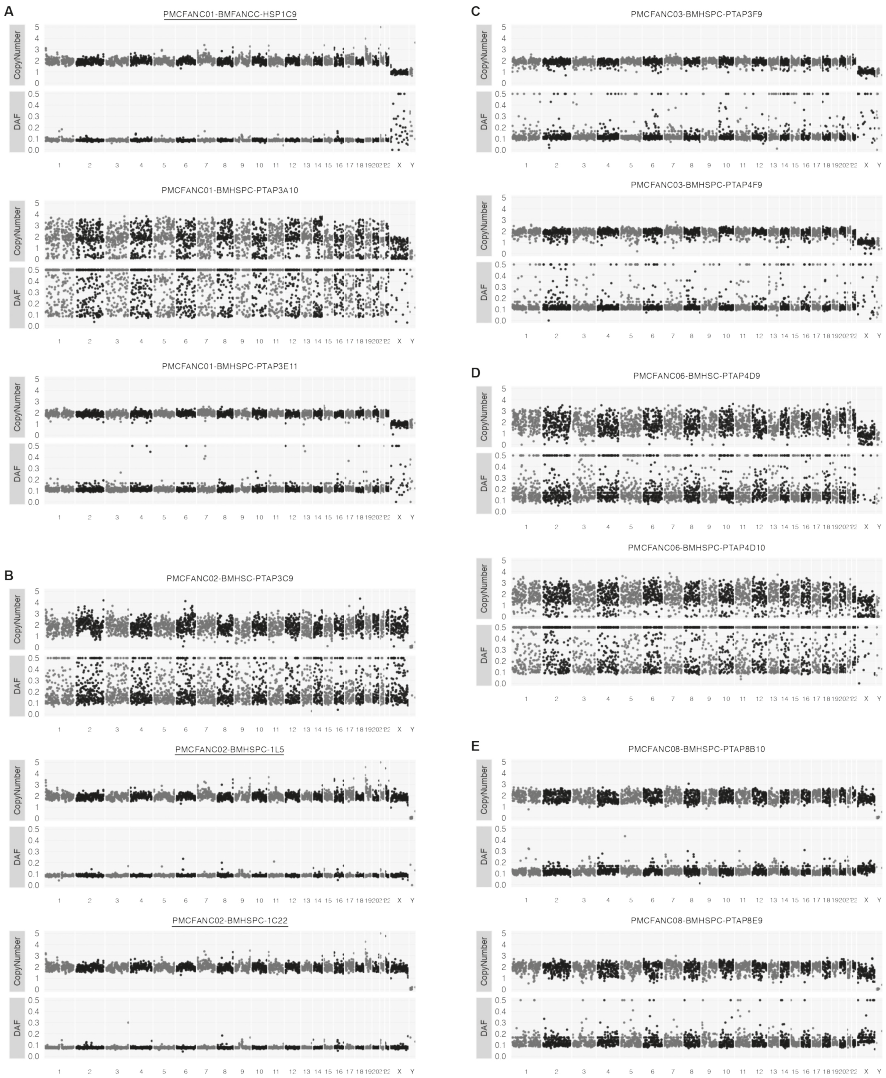


Figure S12. No large chromosomal rearrangements detected in the HSPCs of patients with FA, Related to Figure 5.

(A-E) Copy number and DAF plots (at 1Mb resolution) after PTATO filtering of 12 analyzed HSPCs of 5 patients with FA. There is variability in the PTA quality between the single cells, leading to a lower sensitivity to detect SVs in some samples with relatively low quality (e.g. PMCFANC01-BMHSPC-PTAP3A10 and PMCFANC06-BMHSPC-PTAP4D10). Names of samples that were analyzed by WGS after clonal expansion (instead of PTA) are underlined.

Table S1, S2 and S3. (Scan the QR code to visualize the Table S1, S2 and S3).

References

1. Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science (1979) ***349***, 1483–1489. 10.1126/science.aab4082.
2. Manders, F., van Boxtel, R., and Middelkamp, S. (2021). The Dynamics of Somatic Mutagenesis During Life in Humans. Frontiers in Aging ***2***. 10.3389/fragi.2021.802407.
3. Vijg, J., and Dong, X. (2020). Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. Cell ***182***, 12–23. 10.1016/j.cell.2020.06.024.
4. Ellis, P., Moore, L., Sanders, M.A., Butler, T.M., Brunner, S.F., Lee-Six, H., Osborne, R., Farr, B., Coorens, T.H.H., Lawson, A.R.J., et al. (2021). Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. Nat Protoc ***16***, 841–871. 10.1038/s41596-020-00437-6.
5. Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell ***150***, 264–278. 10.1016/j.cell.2012.06.023.
6. 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. 10.1038/nature19768.
7. Dou, Y., Gold, H.D., Luquette, L.J., and Park, P.J. (2018). Detecting Somatic Mutations in Normal Cells. Trends in Genetics ***34***, 545–557. 10.1016/j.tig.2018.04.003.
8. Ceccaldi, R., Sarangi, P., and D'Andrea, A.D. (2016). The Fanconi anaemia pathway: New players and new functions. Nat Rev Mol Cell Biol ***17***, 337–349. 10.1038/nrm.2016.48.
9. Taylor, A.M.R., Rothblum-Oviatt, C., Ellis, N.A., Hickson, I.D., Meyer, S., Crawford, T.O., Smogorzewska, A., Pietrucha, B., Weemaes, C., and Stewart, G.S. (2019). Chromosome instability syndromes. Nat Rev Dis Primers ***5***, 64. 10.1038/s41572-019-0113-0.
10. Nalepa, G., and Clapp, D.W. (2018). Fanconi anaemia and cancer: An intricate relationship. Nat Rev Cancer ***18***, 168–185. 10.1038/nrc.2017.116.
11. Garaycoechea, J.I., Crossan, G.P., Langevin, F., Daly, M., Arends, M.J., and Patel, K.J. (2012). Genotoxic consequences of endogenous aldehydes on mouse haematopoietic stem cell function. Nature ***489***, 571–575. 10.1038/nature11368.
12. Garaycoechea, J.I., Crossan, G.P., Langevin, F., Mulderrig, L., Louzada, S., Yang, F., Guilbaud, G., Park, N., Roerink, S., Nik-Zainal, S., et al. (2018). Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. Nature ***553***, 171–177. 10.1038/nature25154.
13. Shen, X., Wang, R., Kim, M.J., Hu, Q., Hsu, C.C., Yao, J., Klages-Mundt, N., Tian, Y., Lynn, E., Brewer, T.F., et al. (2020). A Surge of DNA Damage Links Transcriptional Reprogramming and Hematopoietic Deficit in Fanconi Anemia. Mol Cell ***80***, 1013-1024.e6. 10.1016/j.molcel.2020.11.040.
14. Lévy, C., Amirache, F., Girard-Gagnepain, A., Frecha, C., Roman-Rodríguez, F.J., Bernadin, O., Costa, C., Nègre, D., Gutierrez-Guerrero, A., Vranckx, L.S., et al. (2017). Measles virus envelope pseudotyped lentiviral vectors transduce quiescent human HSCs at an efficiency without precedent. Blood Adv ***1***, 2088–2104. 10.1182/bloodadvances.2017007773.
15. Adair, J.E., Chandrasekaran, D., Sghia-Hughes, G., Haworth, K.G., Woolfrey, A.E., Burroughs, L.M., Choi, G.Y., Becker, P.S., and Kiem, H.P. (2018). Novel lineage depletion preserves autologous blood stem cells for gene therapy of fanconi anemia complementation group A. Haematologica ***103***, 1806–1814. 10.3324/haematol.2018.194571.

16. Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. Nat Rev Genet ***17***, 175–188. 10.1038/nrg.2015.16.
17. Gonzalez-Pena, V., Natarajan, S., Xia, Y., Klein, D., Carter, R., Pang, Y., Shaner, B., Annu, K., Putnam, D., Chen, W., et al. (2021). Accurate genomic variant detection in single cells with primary template-directed amplification. Proc Natl Acad Sci U S A ***118***. 10.1073/pnas.2024176118.
18. Luquette, L.J., Miller, M.B., Zhou, Z., Bohrson, C.L., Zhao, Y., Jin, H., Gulhan, D., Ganz, J., Bizzotto, S., Kirkham, S., et al. (2022). Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. Nat Genet ***54***, 1564–1571. 10.1038/s41588-022-01180-2.
19. Bohrson, C.L., Barton, A.R., Lodato, M.A., Rodin, R.E., Luquette, L.J., Viswanadham, V. V, Gulhan, D.C., Cortés-Ciriano, I., Sherman, M.A., Kwon, M., et al. (2019). Linked-read analysis identifies mutations in single-cell DNA-sequencing data. Nat Genet ***51***, 749–754. 10.1038/s41588-019-0366-2.
20. de Kanter, J.K., Peci, F., Bertrums, E., Rosendahl Huber, A., van Leeuwen, A., van Roosmalen, M.J., Manders, F., Verheul, M., Oka, R., Brandsma, A.M., et al. (2021). Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. Cell Stem Cell ***28***, 1726-1739.e6. 10.1016/j.stem.2021.07.012.
21. Luquette, L.J., Bohrson, C.L., Sherman, M.A., and Park, P.J. (2019). Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nat Commun ***10***. 10.1038/s41467-019-11857-8.
22. Gonzalez-Perez, A., Sabarinathan, R., and Lopez-Bigas, N. (2019). Local Determinants of the Mutational Landscape of the Human Genome. Cell ***177***, 101–114. 10.1016/j.cell.2019.02.051.
23. Zou, X., Owusu, M., Harris, R., Jackson, S.P., Loizou, J.I., and Nik-Zainal, S. (2018). Validating the concept of mutational signatures with isogenic cell models. Nat Commun ***9***, 1–16. 10.1038/s41467-018-04052-8.
24. Zou, X., Koh, G.C.C., Nanda, A.S., Degasperi, A., Urgo, K., Roumeliotis, T.I., Agu, C.A., Badja, C., Momen, S., Young, J., et al. (2021). A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. Nat Cancer ***2***, 643–657. 10.1038/s43018-021-00200-0.
25. Drost, J., van Boxtel, R., Blokzijl, F., Mizutani, T., Sasaki, N., Sasselli, V., de Ligt, J., Behjati, S., Grolleman, J.E., van Wezel, T., et al. (2017). Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science (1979) ***358***, 234–238. 10.1126/science.aao3130.
26. 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. 10.1016/j.cell.2019.03.001.
27. 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. 10.1016/j.celrep.2018.11.014.
28. Brandsma, A.M., Bertrums, E.J.M., van Roosmalen, M.J., Hofman, D.A., Oka, R., Verheul, M., Manders, F., Ubels, J., Belderbos, M.E., and van Boxtel, R. (2021). Mutation Signatures of Pediatric Acute Myeloid Leukemia and Normal Blood Progenitors Associated with Differential Patient Outcomes. Blood Cancer Discov ***2***, 484–499. 10.1158/2643-3230.bcd-21-0010.
29. Nik-Zainal, S., van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012). The life history of 21 breast cancers. Cell ***149***, 994–1007. 10.1016/j.cell.2012.04.023.

30. Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell ***149***, 979–993. 10.1016/j.cell.2012.04.024.
31. Sebert, M., Gachet, S., Leblanc, T., Rousseau, A., Bluteau, O., Kim, R., Ben Abdelali, R., Sicre de Fontbrune, F., Maillard, L., Fedronie, C., et al. (2023). Clonal hematopoiesis driven by chromosome 1q/MDM4 trisomy defines a canonical route toward leukemia in Fanconi anemia. Cell Stem Cell ***30***, 153-170.e9. 10.1016/j.stem.2023.01.006.
32. Webster, A.L.H., Sanders, M.A., Patel, K., Dietrich, R., Noonan, R.J., Lach, F.P., White, R.R., Goldfarb, A., Hadi, K., Edwards, M.M., et al. (2022). Genomic signature of Fanconi anaemia DNA repair pathway deficiency in cancer. Nature ***612***, 495–502. 10.1038/s41586-022-05253-4.
33. Mallory, X.F., Edrisi, M., Navin, N., and Nakhleh, L. (2020). Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biol ***21***, 1–22. 10.1186/s13059-020-02119-8.
34. Cameron, D.L., Baber, J., Shale, C., Valle-Inclan, J.E., Besselink, N., van Hoeck, A., Janssen, R., Cuppen, E., Priestley, P., and Papenfuss, A.T. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol ***22***, 1–25. 10.1186/s13059-021-02423-x.
35. 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. 10.1038/s41586-019-1689-y.
36. Xia, Y., Gonzales-Pena, V., Klein, D.J., Luquette, J.J., Puzon, L., Siddiqui, N., Reddy, V., Park, P., Behr, B.R., and Gawad, C. (2021). Genome-wide Disease Screening in Early Human Embryos with Primary Template-Directed Amplification. bioRxiv, 2021.07.06.451077. 10.1101/2021.07.06.451077.
37. Zawistowski, J.S., Salas-González, I., Morozova, T. v, Blackinton, J.G., Tate, T., Arvapalli, D., Velivela, S., Harton, G.L., Marks, J.R., Hwang, E.S., et al. (2022). Unifying genomics and transcriptomics in single cells with ResolveOME amplification chemistry to illuminate oncogenic and drug resistance mechanisms. bioRxiv, 2022.04.29.489440. https://doi.org/10.1101/2022.04.29.489440.
38. Miller, M.B., Huang, A.Y., Kim, J., Zhou, Z., Kirkham, S.L., Maury, E.A., Ziegenfuss, J.S., Reed, H.C., Neil, J.E., Rento, L., et al. (2022). Somatic genomic changes in single Alzheimer's disease neurons. Nature ***604***, 714–722. 10.1038/s41586-022-04640-1.
39. Chen, C.C., Feng, W., Lim, P.X., Kass, E.M., and Jasin, M. (2018). Homology-Directed Repair and the Role of BRCA1, BRCA2, and Related Proteins in Genome Integrity and Cancer. Annu Rev Cancer Biol ***2***, 313–336. 10.1146/annurev-cancerbio-030617-050502.
40. Neveling, K., Endt, D., Hoehn, H., and Schindler, D. (2009). Genotype-phenotype correlations in Fanconi anemia. Mutat Res ***668***, 73–91. 10.1016/j.mrfmmm.2009.05.006.
41. Turajlic, S., Sottoriva, A., Graham, T., and Swanton, C. (2019). Resolving genetic heterogeneity in cancer. Nat Rev Genet ***20***. 10.1038/s41576-019-0114-6.
42. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc ***8***, 2281–2308. 10.1038/nprot.2013.143.
43. Brinkman, E.K., Chen, T., Amendola, M., and Van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Res ***42***, 1–8. 10.1093/nar/gku936.
44. Puschhof, J., Pleguezuelos-Manzano, C., Martinez-Silgado, A., Akkerman, N., Saftien, A., Boot, C., de Waal, A., Beumer, J., Dutta, D., Heo, I., et al. (2021). Intestinal organoid cocultures with microbes. Nat Protoc ***16***, 4633–4649. 10.1038/s41596-021-00589-z.

45. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. 10.1093/bioinformatics/btp324.
46. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. Bioinformatics *31*, 2032–2034. 10.1093/bioinformatics/btv098.
47. Depristo, M.A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491–501. 10.1038/ng.806.
48. Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive. Nucleic Acids Res *39*, D19–D21. 10.1093/nar/gkq1019.
49. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., and Davies, R.M. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*, 1–4. 10.1093/gigascience/giab008.
50. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One *11*, 1–10. 10.1371/journal.pone.0163962.
51. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat Biotechnol *35*, 316–319. 10.1038/nbt.3820.
52. Shale, C., Cameron, D.L., Baber, J., Wong, M., Cowley, M.J., Papenfuss, A.T., Cuppen, E., and Priestley, P. (2022). Unscrambling cancer genomes via integrated analysis of structural variation and copy number. Cell Genomics *2*, 100112. 10.1016/j.xgen.2022.100112.
53. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, and D. (2002). The Human Genome Browser at UCSC. Genome Res *12*, 996–1006. 10.1101/gr.229102.
54. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res *48*, D682–D688. 10.1093/nar/gkz966.
55. Delaneau, O., Zagury, J.F., Robinson, M.R., Marchini, J.L., and Dermitzakis, E.T. (2019). Accurate, scalable and integrative haplotype estimation. Nat Commun *10*, 24–29. 10.1038/s41467-019-13225-y.
56. Wickham (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).
57. Obenchain, V., Lawrence, M., Carey, V., Gogarten, S., Shannon, P., and Morgan, M. (2014). VariantAnnotation: A Bioconductor package for exploration and annotation of genetic variants. Bioinformatics *30*, 2076–2078. 10.1093/bioinformatics/btu168.
58. Manders, F., Brandsma, A.M., de Kanter, J., Verheul, M., Oka, R., van Roosmalen, M.J., van der Roest, B., van Hoeck, A., Cuppen, E., and van Boxtel, R. (2022). MutationalPatterns: the one stop shop for the analysis of mutational processes. BMC Genomics *23*, 1–18. 10.1186/s12864-022-08357-3.
59. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.
60. Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., and Hall, I.M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods *12*, 966–968. 10.1038/nmeth.3505.
61. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101. 10.1038/s41586-020-1943-3.

62. Lüdecke, D. (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. J Open Source Softw *3*, 772. 10.21105/joss.00772.
63. Nilsen, G., Liestøl, K., Van Loo, P., Moen Vollan, H.K., Eide, M.B., Rueda, O.M., Chin, S.-F., Russell, R., Baumbusch, L.O., Caldas, C., et al. (2012). Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics *13*, 591. 10.1186/1471-2164-13-591.
64. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639–1645. 10.1101/gr.092759.109.
65. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24–26. 10.1038/nbt.1754.

GENERAL DISCUSSION

7

Flavia Peci[1,2]

[1]Princess Máxima Center for Pediatric Oncology, Utrecht, 3584 CS, The Netherlands
[2]Oncode Institute, Jaarbeursplein 6, 3521 AL Utrecht, The Netherlands

INTRODUCTION

For nearly half a century, hematopoietic stem cell transplantation (HSCT) has served as an essential treatment modality for a wide range of malignant and non-malignant hematological disorders globally. However, our understanding of human hematopoiesis regeneration in the recipient upon transplant and the factors that influence the long-term maintenance of hematopoietic stem cells throughout life remains unclear. This thesis presents a thorough investigation into the mutational effects of HSCT, which is performed on thousands of patients annually. It is estimated that the number of HSCT survivors will reach 500,000 in the US by 2030. Despite being a life-saving procedure, individuals undergoing transplantation still encounter various life-threatening side effects high rate of late mortality risk[1].

TREATMENT EFFECTS OF HSCT ON THE RECIPIENT BONE MARROW NICE

Patients who undergo transplantation are at an increased risk of developing secondary malignancies, clonal hematopoiesis, and bacterial, viral, and fungal infections, although the cause of this increased risk is not yet clear. The mechanisms behind these complications are not fully understood. One factor that needs to be considered, is the role of the bone marrow niche in providing an instructed microenvironment for the donor transplanted hematopoietic stem and progenitor cells (HSPCs) to engraft and regenerate the recipient blood system. As discussed in **chapter 2**, pre-transplant conditioning regimens, which consist of a combination of chemotherapeutic drugs (busulfan, clofarabine, fludarabine), monoclonal antibody therapy, and/or total body irradiation, are likely to negatively affect the engraftment of stem cells and hamper the success of HSCT because of the damage caused to non-hematopoietic niche cells[3]. Indeed, toxicity in BM endothelial cells (ECs) caused by irradiation and high-dose cyclophosphamide or busulfan has been associated with several EC-related disorders, from veno-occlusive disease to sinusoidal obstruction syndrome[4]. Protecting ECs from conditioning-related damage could preserve their cytokine production, which is essential to boost hematopoietic regeneration after transplant[3]. Administration of pigment endothelial-derived factor (PEDF), defibrotide, and N-acetyl-L-cysteine (NAC) could prevent such damage and complement current treatment regimens[3]. Another cell type particularly affected in the recipient BM niche is the mesenchymal stromal cell (MSC). MSCs may be considered the backbone of the BM niche because of their intrinsic capacity to differentiate into various cell types that support HSPCs, such as bone, cartilage, adipose tissue, tendon and muscle[5]. Recent work has shown

that pre-transplant regimens damage MSCs by inducing DNA double-strand breaks, altering gene expression, and skewing their differentiation towards an osteogenic fate[6]. Therefore, by combining conditioning treatments with therapies that protect the non-hematopoietic niche fraction, it may be possible to limit cellular damage to the niche and improve the long-term success of HSCT. However, the development of these therapies is currently hampered by the lack of a proper human model that functionally and transcriptionally recapitulates the BM niche environment. In this regard, the recent development of BM niche organoids[7] will likely provide a useful tool for the investigation of hematopoiesis in health and disease. Although still in the preliminary stages, human BM niche organoids might also offer an ideal platform to study hematopoietic malignancies, drug development and screening[7].

ANTIVIRAL NUCLEOSIDE ANALOG TREATMENT IN HSCT

During the HSCT procedure, immunosuppression is required to allow successful stem cells engraftment of donor cells in the host and prevent alloreactive conditions, such as GvHD[8]. However, even though this immunosuppression is temporary, it renders the HSCT recipient vulnerable to a large class to opportunistic bacterial, fungal, and viral infections. The risk of specific viral infections depends on a series of factors, namely serostatus of both the donor and recipient, as well as the type of transplant and stem cell source[1]. One of the most common and clinically significant viral agents in the transplantation setting is Cytomegalovirus (CMV), which re-activation occurs in ~60-70% of CMV-seropositive transplanted patients, and 20-30% of CMV-seronegative recipients transplanted from CMV-seropositive donors[9]. CMV can be treated with letermovir, a recently approved compound, but also with cidofovir, foscarnet and (val-)ganciclovir (GCV). Whilst our research focus was originally to investigate HSCT-associated mutagenesis in transplanted stem cells, we discovered that GCV treatment resulted in an increased number of SBS in the transplanted HSPCs. In-depth mutational analyses revealed that GCV induced a specific mutational signature, which we could validate by *in vitro* exposing UCB-derived HSPCs to the compound. The results this study can be found in **chapter 3**. For this, we used a genotoxicity assay described in **chapter 4.** Furthermore, given that we and others found GCV to be highly mutagenic and potentially carcinogenic[10,11], we aimed to screen more antiviral nucleoside analogs (NAs) used in the treatment of viral infections. In **chapter 5**, we used our genotoxicity assay to test the mutational consequences of a collection of NAs when exposed to human HSPCs. From the list, we concluded that most antiviral NAs were not mutagenic. However, 6 out of 16 compounds used in the

screening resulted in increased number of mutations. Among these mutagenic treatments, we found treatments for HIV (AZT, DDC), Sars-CoV-2 (MOV), HSV (PCV, BVD) and GCV. Little is currently known about the long-term clinical impact of these mutations in transplanted HSPCs and human in general and what outcome this might lead to. A hypothesis is that using drugs that lead to enhanced mutagenesis and cause genetic drivers in immunosuppressed/immunocompromised patients may elevate the risk of second malignancies later in life. However, for now the long-term clinical impact and side effects of these drugs is outweighed by their therapeutic benefit. Further research is needed to understand the long-term clinical impact of mutations in transplanted HSPCs and their implications for human health. This research should focus on assessing the outcomes and potential risks associated with these mutations, particularly in immunosuppressed or immunocompromised patients and contribute to the formulation of safer antiviral drugs. Furthermore, conducting clinical studies and trials can help evaluate the effectiveness and safety of drugs that lead to enhanced mutagenesis. These trials should specifically investigate the risk of developing second malignancies later in life upon using such drugs. Additionally, considering the long-term clinical impact and potential treatment related risks, regulatory bodies should develop guidelines and recommendations for healthcare professionals. For instance, the use of a different compound with reduced mutagenic risk should be preferred.

GENOME ANALYSIS OF HSPCs BY SINGLE CELL SEQUENCING

Sequencing a nucleic acid molecule, whether it's DNA or RNA, allows for deciphering the genetic information that is crucial for life. The groundbreaking discovery of the double-helical DNA structure by Watson and Crick laid the foundation for the advancement of sequencing technologies[12]. These technologies can be traced back to the 1970s when pioneering researchers like Walter Fiers, Frederick Sanger, and Ray Wu made significant strides in deciphering the DNA code through sequencing methods[13]. Several decades later, thanks to the rapid advancements in methods and technologies, coupled with reduced costs and time, single-cell genomics has greatly improved our understanding of tissue composition in both health and disease[14,15,16,17,18]. This approach, particularly in the context of blood analysis, enables the determination of cellular heterogeneity within human tissues through the examination of somatic mutations and transcriptome[19,20]. To study the clonality and dynamics of blood tissue, an *in vitro* clonal expansion of HSPCs has to be employed to obtain sufficient DNA for whole-genome sequencing (WGS) analysis[21]. WGS allows for investigating the entire genome, including non-

coding regions and all types of mutations. This is in contrast to whole exome sequencing (WES), which covers approximately 2% of the human genome but includes 85% of known pathogenic variants, or targeted sequencing panels that focus on genes associated with specific diseases or disease groups[22]. Research on single HSPC genomes has facilitated tracking the initiation and development of clonal hematopoiesis (CH) over time in general hematopoiesis dynamics across human lifespan[23,24] and transplanted patients[25]. These studies have shown us that donor-engrafted CH, i.e., transfer of clones from donor to recipient occurred frequently (in about 12% of transplantations) and that transplantations from older donors is more likely to result in donor-engrafted CH[25,26]. Additionally, single cell WGS data can be used in the analysis of mutational signatures and phylogenetic reconstruction allows us to determine the cell of origin of cancer. These analyses can provide useful insights when examining the effects of chemotherapy and nucleoside antiviral treatments on hematopoiesis throughout a patient's lifespan or following a transplant[10,27]. It is important to note that numerous studies have confirmed that such treatments leave specific genomic scars in healthy cells, potentially leading to the development of second cancers[10,28,27]. I envision that in the future, there will be a drive in research to reduce the amount of DNA required for single-cell sequencing and to perform multi-omics analysis from a single cell. For now, whole genome amplification (WGA) is required for WGS. For single HSPCs WGA can be performed in culture as these cells retain self-renewal capacity and give rise to genomically identical cells. The starting DNA content of one cell is therefore exponentially multiplied. However, when *in vitro* clonal expansion of cells is not possible from a single cell because of their lost capacity to clonally expand (like in Fanconi Anemia disorder), other methods are needed to obtain enough DNA from a single patient blood stem cell. In the following paragraph, we will summaries important aspects/findings on the Fanconi Anemia (FA) HSCs genome stability using a novel method that utilizes WGA prior to WGS.

THE CHALLENGE OF FANCONI ANEMIA STEM CELLS

HSCT is a highly effective procedure used to treat a range of diseases, including progressive bone marrow failure (BMF) which occurs in the inherited genetic disorder named Fanconi Anemia (FA). Of note, besides the multiple and severe clinical manifestations, the pathophysiology of FA BMF still remain elusive, and HSCT remains today the only curative treatment[29]. Recent studies investigating the underlying molecular mechanisms of BMF in FA have identified the role of endogenous aldehyde-induced toxicity and/or DNA damage-induced p53 activation. These processes ultimately lead to the depletion of HSPCs[30].

Notably, the development of BMF has been observed in FA mice that also carry mutations in aldehyde dehydrogenase 2 (Aldh2), providing valuable models for studying hematopoiesis in FA[30,31]. Interestingly, individuals with mutations in ALDH2, similar to the FA mice, demonstrate an accelerated progression of BMF[32]. Furthermore, FA is characterized by constitutive genomic instability, which creates a favorable environment for clonal evolution and tumor progression. Consequently, patients with FA have a significantly higher risk of developing cancer[33]. Most research on the mutagenic processes in FA HSPCs have been performed using mouse models, due to the delicate genomic nature of the FA HSPCs cells, which has hindered the study of FA HSPCs genome on a single cell level. Whilst common single cell WGS methods rely on *in vitro* WGA by clonal expansion of the stem cells of interest to obtain enough DNA, *in vitro* clonal expansion of FA HSPCs remains challenging. Of note, FA HSPCs are compromised by unresolved interstrand cross links, chromosomal breaks and genome instability, which results in high level of DNA damage and activated p53 pathway which leads to apoptosis[30]. To overcome this limitation, we employed a novel WGA method, called primary template-directed amplification (PTA), to obtain sufficient DNA for WGS analysis of a single HSPC. Because PTA introduces mutational artefacts in the WGS data of the cell, an in house bioinformatics pipeline called PTA Analysis Tool (PTATO) has been developed to filter out these artefacts and obtain an accurate estimate of the FA HSPCs mutation burden[34]. With the aid of PTATO, described in **chapter 6**, we analyzed the genomes of individual FA HSPCs. Surprisingly, our analysis revealed that the number of somatic base substitutions in FA HSPCs is comparable to that of healthy donor-derived HSPCs. However, we observed an increased frequency of deletions, indicating significant genomic instability[34]. These findings are consistent with previous *in vivo* and *in vitro* studies[31,35]. The knowledge gained from studying somatic mutations in single HSPCs of FA patients not only enhances our understanding of the mechanisms underlying this specific condition, but also holds implications for genes implicated in the FA pathway and associated with other diseases. For example, dysregulation of FANCD2/BRCA2 gene, implicated in FA pathway, also plays a major role in breast and ovarian cancer[36,37].

GENETIC SAFETY OF HSCT

As the long-term survival of HSCT patients increased over the past decades[38,39] the need to ensure the genetic safety of the transplantation procedure, particularly in pediatric patients, has sparked great interest. This interest stems from two primary reasons. Firstly, there is a need to address potential genetic risks and complications that may emerge years after the transplant. These

risks can include late effects of conditioning treatments or the phenomenon of donor-derived clonal hematopoiesis and hematological malignancies. Secondly, certain genetic disorders or predispositions may not be evident until later in life. Consequently, it becomes crucial to assess the genetic safety of the transplantation procedure to identify and mitigate the risk of late-onset genetic disorders in HSCT patients. Besides, the compatibility between a donor and recipient in HSCT depends on several important factors, including HLA mismatch, viral status, willingness to donate, and age. Among these factors, donor age has emerged as a crucial non-Human Leukocyte Antigen (HLA) characteristic in unrelated donor HSCT. It is worth noting that patients can receive stem cells for transplantation from donors up to the age of 40. However, there has been a rise in the utilization of haplo-transplantations, where graft material is derived from a parent. In such cases, the parents are inevitably several years or even decades older than the recipient. Nevertheless, the mutational consequences associated with aging and the assessment of mutational quality in stem cells have only recently begun to receive attention and investigation[40]. Age-related somatic mutation accumulation can contribute to cancer development[41,42,43]. Somatic mutations are known to accumulate in a tissue-specific manner over time[43]. In healthy human blood stem cells, somatic mutation accumulates at the rate of ~15-17 mutations per year[21]. However, this accumulation rate may be perturbed in case of emergency hematopoiesis, such as after HSCT. Indeed, replication stress triggered by transplantation was shown to increase intracellular level of reactive oxygen species, which in turn caused DNA damage[44]. Currently, it is still debated how many of the transplanted stem cells can engraft and regenerate the hematopoietic system as well as how many cell divisions stem cells undergo to suffice the hematopoietic regenerative demand. In **chapter 3**, we aimed to catalogue all the somatic mutations in HSPCs collected before and after HSPCs in pediatric patients. By applying single cell WGS to HSPCs collected from 9 pediatric donor and recipient matched siblings' pairs, before and up to 2 years after HSCT, we measured the somatic mutation burden and mutational profiles. Our cohort was composed by patients who have been transplanted with an indication of hematological malignancies and had successful transplant. Also, the cohort included patient transplanted with different stem cells sources, namely cord blood, in which stem cells are collected from umbilical cord blood donor (cord blood transplantation), bone marrow stem cells donated by a HLA-compatible parent (haploidentical stem cell transplantation) or donated by an HLA-compatible sibling (siblings matched transplantation). Interestingly, by looking at these patients up to 2 years post-transplant, we did not observe increased

mutation load in most of the transplanted HSPCs analyzed, compared to their matched sibling's donor HSPCs collected at the same time. Therefore, no increased mutational ageing was detected[1]. In contrast, a recent study from Campbell P et al. on clonal dynamics of recipients after 9-31 years post HSCT, reported increased ageing in transplanted HSPCs compared to their matched donors (recipient HSPCs had ~23 excess mutations equivalent to 1.5 years of normal ageing). This increase was equivalent to ~10-15 years of additional ageing[2]. Although in our cohort we collected material from patients up to 2 years post-HSCT and did not find any overall evidence of increased mutation load in HSPCs, 2 out of 9 patients of our cohort showed an increased number of single base substitutions (SBS) and presence of increased number of C>A mutations in 100% and 43% of recipient HSPCs. By examining the mutational signature analysis, we found that the clones exhibiting increased mutagenesis displayed a significant presence of a novel mutational signature known as SBSA. This particular signature was characterized by the C>A mutation occurring at the CpG dinucleotide context. Further investigation led us to attribute this mutational signature to the antiviral drug GCV. GCV is a guanosine analog commonly employed as the gold standard treatment for post-HSCT CMV reactivation in clinical settings. Furthermore, by applying a machine learning approach in pan-cancer genome datasets, we found the presence of SBSA in a treatment-related AMLs and a dataset of metastatic cancers. We calculated the chance of SBSA of causing driver mutations in 3 different type of cancers and found that the signature had high contribution of inducing driver mutations. Despite its mutagenicity and demonstrated carcinogenic potential, GCV is still used as first line treatment for the management of CMV, which remains the most clinically significant infection in immunosuppressed patient's groups such as HSCT patients, solid organ transplanted patients or those with sever immune deficiency or HIV/AIDS[45–47]. CMV clinical infections can also be treated with Foscarnet, a pyrophosphate analog which inhibits the viral polymerase activity. However, it's use is currently hampered by nephrotoxicity limitations[48]. A new and recent approach to CMV re-activation post-transplant is offered by letermovir, which was approved in 2017[49]. Letermovir, which is an antiviral agent that targets DNA terminal transferase complex of the CMV, has been employed for the prophylaxis in HSCT transplanted recipients and lowered the risk of clinically significant CMV infection with low grade side effects[50].

CONCLUDING REMARKS

HSCT stands out as one of the most captivating and primary cellular therapies available today. It is a suitable treatment for a wide range of malignant and non-malignant hematopoietic disorders in both adult and pediatric patients. Despite its established use and acclaim, there is still room for optimization to enhance long term engraftment of HSPCs and hematopoietic reconstitution, patients' survival and quality of life. One critical area for improvement is reducing the hematologic malignancy relapse rate, a significant drawback following transplantation[51,52]. This thesis highlights the potential of single cell sequencing technologies in HSCT patients to yield better outcomes for cancer survivors. By enabling stricter monitoring of disease progression and recovery status over time post-HSCT, these technologies can aid in understanding the biology behind clonal hematopoiesis, relapse, and potentially other pathologies related to the life of patients after HSCT. Moreover, studying hematopoietic regeneration in humans can deepen our understanding of HSPCs biology. This knowledge can contribute to the development of future customized treatments, such as the transplantation of ex-vivo quality-controlled expanded HSCs[53,54], customized graft manipulation, and gene therapy for inherited genetic disorders like FA[55,56]. These approaches will provide a platform to deliver specific and genomically safe cell types tailored to the recipient's needs, overcoming the shortage of donor material as well as removing the genetic risks deriving from the latter[53]. Furthermore, recent population-wide or geographic blood and tumor sequencing screenings have unveiled that genome analysis, such as studying mutation types and signatures, can offer crucial insights into the etiology of cancers[57] and conditions such as clonal hematopoiesis[58]. This knowledge can drive collective efforts between the scientific community and governments to enhance public health[57,58,59]. Examples include eliminating known carcinogens from food or the environment and lowering the age for cancer screening in specific tumors[60].

ACKNOWLEDGEMENTS

I extend my gratitude to Ruben van Boxtel and Mirjam Belderbos for providing valuable feedback on this work.

REFERENCES

1. Bhatia, S. *et al.* Trends in Late Mortality and Life Expectancy After Allogeneic Blood or Marrow Transplantation Over 4 Decades: A Blood or Marrow Transplant Survivor Study Report. *JAMA Oncol.* **7**, 1626 (2021).
2. Bhatia, S. *et al.* Trends in Late Mortality and Life Expectancy After Autologous Blood or Marrow Transplantation Over Three Decades: A BMTSS Report. *J. Clin. Oncol.* **40**, 1991–2003 (2022).
3. Peci, F. *et al.* The cellular composition and function of the bone marrow niche after allogeneic hematopoietic cell transplantation. *Bone Marrow Transplant.* **57**, 1357–1364 (2022).
4. Hildebrandt, G. C. & Chao, N. Endothelial cell function and endothelial-related disorders following haematopoietic cell transplantation. *Br. J. Haematol.* **190**, 508–519 (2020).
5. Gao, Q. *et al.* Bone Marrow Mesenchymal Stromal Cells: Identification, Classification, and Differentiation. *Front. Cell Dev. Biol.* **9**, 787118 (2022).
6. Preciado, S. *et al.* Mesenchymal Stromal Cell Irradiation Interferes with the Adipogenic/Osteogenic Differentiation Balance and Improves Their Hematopoietic-Supporting Ability. *Biol. Blood Marrow Transplant.* **24**, 443–451 (2018).
7. Khan, A. O. *et al.* Human Bone Marrow Organoids for Disease Modeling, Discovery, and Validation of Therapeutic Targets in Hematologic Malignancies. *Cancer Discov.* **13**, 364–385 (2023).
8. Gyurkocza, B. & Sandmaier, B. M. Conditioning regimens for hematopoietic cell transplantation: one size does not fit all. *Blood* **124**, 344–353 (2014).
9. Einsele, H., Ljungman, P. & Boeckh, M. How I treat CMV reactivation after allogeneic hematopoietic stem cell transplantation. *Blood* **135**, 1619–1629 (2020).
10. de Kanter, J. K. *et al.* Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. *Cell Stem Cell* **28**, 1726-1739.e6 (2021).
11. Fang, H. *et al.* Ganciclovir-induced mutations are present in a diverse spectrum of post-transplant malignancies. *Genome Med.* **14**, 124 (2022).
12. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
13. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).
14. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
15. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* (2023) doi:10.1038/s41580-023-00615-w.
16. Jia, Q., Chu, H., Jin, Z., Long, H. & Zhu, B. High-throughput single-cell sequencing in cancer research. *Signal Transduct. Target. Ther.* **7**, 145 (2022).
17. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
18. García-Nieto, P. E., Morrison, A. J. & Fraser, H. B. The somatic mutation landscape of the human body. *Genome Biol.* **20**, 298 (2019).
19. Pellin, D. *et al.* A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* **10**, 2395 (2019).
20. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).

21. Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep.* **25**, 2308-2316.e4 (2018).
22. Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human Genome Sequencing in Health and Disease. *Annu. Rev. Med.* **63**, 35–61 (2012).
23. Fabre, M. A. *et al.* The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335–342 (2022).
24. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
25. Boettcher, S. *et al.* Clonal hematopoiesis in donors and long-term survivors of related allogeneic hematopoietic stem cell transplantation. *Blood* **135**, 1548–1559 (2020).
26. Nawas, M. T. *et al.* The clinical implications of clonal hematopoiesis in hematopoietic cell transplantation. *Blood Rev.* **46**, 100744 (2021).
27. Bertrums, E. J. M. *et al.* Elevated Mutational Age in Blood of Children Treated for Cancer Contributes to Therapy-Related Myeloid Neoplasms. *Cancer Discov.* OF1–OF14 (2022) doi:10.1158/2159-8290.CD-22-0120.
28. Diamond, B. *et al. Chemotherapy Signatures Map Evolution of Therapy-Related Myeloid Neoplasms*. http://biorxiv.org/lookup/doi/10.1101/2022.04.26.489507 (2022) doi:10.1101/2022.04.26.489507.
29. Ebens, C. L., MacMillan, M. L. & Wagner, J. E. Hematopoietic cell transplantation in Fanconi anemia: current evidence, challenges and recommendations. *Expert Rev. Hematol.* **10**, 81–97 (2017).
30. Ceccaldi, R. *et al.* Bone Marrow Failure in Fanconi Anemia Is Triggered by an Exacerbated p53/p21 DNA Damage Response that Impairs Hematopoietic Stem and Progenitor Cells. *Cell Stem Cell* **11**, 36–49 (2012).
31. Garaycoechea, J. I. *et al.* Alcohol and endogenous aldehydes damage chromosomes and mutate stem cells. *Nature* **553**, 171–177 (2018).
32. Hira, A. *et al.* Variant ALDH2 is associated with accelerated progression of bone marrow failure in Japanese Fanconi anemia patients. *Blood* **122**, 3206–3209 (2013).
33. Quentin, S. *et al.* Myelodysplasia and leukemia of Fanconi anemia are associated with a specific pattern of genomic abnormalities that includes cryptic RUNX1/AML1 lesions. *Blood* **117**, e161–e170 (2011).
34. Middelkamp, S. *et al. Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox*. http://biorxiv.org/lookup/doi/10.1101/2023.02.15.528636 (2023) doi:10.1101/2023.02.15.528636.
35. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1744 (2018).
36. Mani, C. *et al.* Hedgehog/GLI1 Transcriptionally Regulates FANCD2 in Ovarian Tumor Cells: Its Inhibition Induces HR-Deficiency and Synergistic Lethality with PARP Inhibition. *Neoplasia* **23**, 1002–1015 (2021).
37. Rudland, P. S. *et al.* Significance of the Fanconi Anemia FANCD2 Protein in Sporadic and Metastatic Human Breast Cancer. *Am. J. Pathol.* **176**, 2935–2947 (2010).
38. Penack, O. *et al.* How much has allogeneic stem cell transplant–related mortality improved since the 1980s? A retrospective analysis from the EBMT. *Blood Adv.* **4**, 6283–6290 (2020).
39. Kliman, D. *et al.* Hematopoietic Stem Cell Transplant Recipients Surviving at Least 2 Years from Transplant Have Survival Rates Approaching Population Levels in the Modern Era of Transplantation. *Biol. Blood Marrow Transplant.* **26**, 1711–1718 (2020).

40. Wang, Y. *et al.* Donor and recipient age, gender and ABO incompatibility regardless of donor source: validated criteria for donor selection for haematopoietic transplants. *Leukemia* **32**, 492–498 (2018).
41. Risques, R. A. & Kennedy, S. R. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLOS Genet.* **14**, e1007108 (2018).
42. Hernando, B. *et al.* The effect of age on the acquisition and selection of cancer driver mutations in sun-exposed normal skin. *Ann. Oncol.* **32**, 412–421 (2021).
43. Manders, F., van Boxtel, R. & Middelkamp, S. The Dynamics of Somatic Mutagenesis During Life in Humans. *Front. Aging* **2**, 802407 (2021).
44. Yahata, T. *et al.* Accumulation of oxidative DNA damage restricts the self-renewal capacity of human hematopoietic stem cells. *Blood* **118**, 2941–2950 (2011).
45. Jakharia, N., Howard, D. & Riedel, D. J. CMV Infection in Hematopoietic Stem Cell Transplantation: Prevention and Treatment Strategies. *Curr. Treat. Options Infect. Dis.* **13**, 123–140 (2021).
46. Lumbreras, C. *et al.* Cytomegalovirus infection in solid organ transplant recipients. *Clin. Microbiol. Infect.* **20**, 19–26 (2014).
47. Murray, J. *et al.* Treating HIV-associated cytomegalovirus retinitis with oral valganciclovir and intra-ocular ganciclovir by primary HIV clinicians in southern Myanmar: a retrospective analysis of routinely collected data. *BMC Infect. Dis.* **20**, 842 (2020).
48. Deray, G. *et al.* Foscarnet Nephrotoxicity: Mechanism, Incidence and Prevention. *Am. J. Nephrol.* **9**, 316–321 (1989).
49. Foolad, F., Aitken, S. L. & Chemaly, R. F. Letermovir for the prevention of cytomegalovirus infection in adult cytomegalovirus-seropositive hematopoietic stem cell transplant recipients. *Expert Rev. Clin. Pharmacol.* **11**, 931–941 (2018).
50. Marty, F. M. *et al.* Letermovir Prophylaxis for Cytomegalovirus in Hematopoietic-Cell Transplantation. *N. Engl. J. Med.* **377**, 2433–2444 (2017).
51. Webster, J. A., Luznik, L. & Gojo, I. Treatment of AML Relapse After Allo-HCT. *Front. Oncol.* **11**, 812207 (2021).
52. Dietz, A. C. & Wayne, A. S. Cells to prevent/treat relapse following allogeneic stem cell transplantation. *Hematology* **2017**, 708–715 (2017).
53. Wang, Y. & Sugimura, R. Ex vivo expansion of hematopoietic stem cells. *Exp. Cell Res.* **427**, 113599 (2023).
54. Wilkinson, A. C. *et al.* Long-term ex vivo haematopoietic-stem-cell expansion allows nonconditioned transplantation. *Nature* **571**, 117–121 (2019).
55. Río, P. *et al.* Successful engraftment of gene-corrected hematopoietic stem cells in non-conditioned patients with Fanconi anemia. *Nat. Med.* **25**, 1396–1401 (2019).
56. Siegner, S. M. *et al.* Adenine base editing efficiently restores the function of Fanconi anemia hematopoietic stem and progenitor cells. *Nat. Commun.* **13**, 6900 (2022).
57. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
58. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nat. Genet.* **54**, 1155–1166 (2022).
59. Degasperi, A. *et al.* Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science* **376**, abl9283 (2022).
60. Rosendahl Huber, A., Van Hoeck, A. & Van Boxtel, R. The Mutagenic Impact of Environmental Exposures in Human Cells and Cancer: Imprints Through Time. *Front. Genet.* **12**, 760039 (2021).

APPENDICES A

NEDERLANDSE SAMENVATTING

Hematopoietische stamceltransplantatie (HSCT) is een levensreddende behandeling voor een breed scala aan aandoeningen, zoals erfelijke genetische aandoeningen als Fanconi Anemie, immunologische aandoeningen en voornamelijk hematologische maligniteiten. Hoewel HSCT de overlevingskansen van patiënten in de afgelopen 40 jaar wereldwijd heeft verbeterd, zijn er nog steeds uitdagingen en bijwerkingen verbonden aan de procedure. Hieronder vallen gebeurtenissen na de transplantatie, zoals de ontwikkeling van acute en chronische graft-versus-hostziekte (GvHD), afstoting van de transplantatie, infecties, klonale hematopoëse en een lagere levensverwachting.

De HSCT-behandeling wordt gekenmerkt door gedefinieerde stappen. Voorafgaand aan de transplantatie dag ondergaan patiënten een myeloablatieve conditionering, een combinatie van hoge dosis chemotherapie en totale lichaamsbestraling. Dit roeit de ziekte uit, creëert ruimte voor de toekomstige donorstamcellen om zich in te nestelen en vermindert de immuunrespons van de ontvanger. Omdat deze behandeling korte- en lange termijn toxiciteit voor de cellen van het beenmerg kan veroorzaken, wat de functie van hematopoëtische stamcellen (HSC) en hematopoëtische regeneratie bij de ontvanger kan belemmeren, is verder onderzoek nodig om transplantatieregimes te optimaliseren en de toxiciteit te verminderen. In **hoofdstuk 2** bieden we een uitgebreid overzicht van het effect van myeloablatieve pre-transplantatie conditionering op de cellulaire componenten van het beenmerg van de ontvanger en beschrijven we hoe dit de hematopoëtische regeneratie kan belemmeren.

HSC's die worden gebruikt voor transplantatiedoeleinden kunnen uit verschillende bronnen worden geïsoleerd, zoals beenmerg, perifeer bloed en navelstrengbloed. Onlangs kwam de leeftijd van de donoren naar voren als een van de belangrijkste niet-HLA-gerelateerde factoren voor een succesvolle transplantatie. Dit komt omdat somatische mutaties in HSC's in de loop van de tijd lineair toenemen, met andere woorden: hoe ouder de donor is, hoe meer mutaties zich al in de transplanteerbare stamcellen hebben opgehoopt. Dit proces wordt ook veroudering genoemd, en is een van de redenen waarom er een leeftijdsgrens is om bloed te doneren. Hoewel sommige mutaties geen invloed hebben op het fenotype, kunnen andere een proliferatief voordeel bieden en in de loop van de tijd de bloedklonaliteit bij de ontvanger

beïnvloeden, wat het risico op klonale hematopoëse verhoogt (een aandoening die gepaard gaat met de ontwikkeling van hematologische maligniteiten).

In het afgelopen decennium hebben de ontwikkelingen in next-generation sequencing (NGS) technologieën het begrip van mutatie-accumulatie in het bloed en andere weefsels verbeterd. Single-cell whole-genome sequencing (WGS) is onlangs naar voren gekomen als een essentiële techniek om het menselijke bloedsysteem te bestuderen. Deze methode maakt de detectie van mutaties, chromosomale afwijkingen en de reconstructie van fylogenetische stambomen mogelijk door somatische mutaties te gebruiken als unieke cellulaire streepjescodes. Door single-cell WGS toe te passen op de studie van hematopoëtische regeneratie na transplantatie kunnen we biologische vragen aanpakken die nog steeds onbeantwoord zijn in het veld, zoals: i) Wat het lange termijn-implanteerbare potentieel en bijdrage van getransplanteerde stamcellen is, en ii) Hoeveel HSC's bijdragen aan het herstel van het hematopoëtische systeem van de ontvanger.

Zoals eerder vermeld, hopen somatische mutaties zich in de loop van de tijd op in het DNA van HSC's. Over het algemeen treden mutaties op door een combinatie van extracellulaire en intracellulaire processen. Er stapelen bijvoorbeeld mutaties op bij elke celdeling en na blootstelling van DNA aan intrinsieke genotoxische stress in cellen. Bovendien kunnen milieublootstelling aan Uv-licht en de inname van carcinogene stoffen zoals tabak of aflatoxine mutaties induceren. Mutaties kunnen ook worden gebruikt om de processen te identificeren die mutagenese en carcinogenese hebben veroorzaakt. Voor dit doel wordt een signatuuranalyse gebruikt als methode om de etiologie van mutaties in verschillende weefsels te onderzoeken. Omdat signaturen bestaan voor alle soorten mutaties die aanwezig zijn in genomische sequenties, is het mogelijk om ze te classificeren op basis van hun kenmerkende eigenschappen, zoals type en voorkomen. Dit resulteert in zeer specifieke patronen die de activiteit van individuele mutagene processen weerspiegelen. De toepassing van mutatiesignaturen (MS) -analyse op de studie van het HSC-genoom onthulde de processen die ten grondslag liggen aan de levenslange accumulatie van mutaties in HSPC's, die werden weerspiegeld door SBS1, SBS5 en het HSPC-signatuur. SBS1, SBS5, en HSPC zijn namelijk klokachtige signaturen en weerspiegelen het verouderingsproces van stamcellen in het hematopoëtische compartiment.

Om het effect van transplantatie op het genoom van HSC's te onderzoeken, hebben we MS-analyse toegepast op getransplanteerde patiënten in ons instituut. We hebben in dit cohort geen verhoogde mutatie gerelateerde veroudering waargenomen.

HSCT-patiënten lopen ook een verhoogd risico op het ontwikkelen van infecties als gevolg van onderdrukking van het immuunsysteem tijdens de behandeling. Antivirale nucleoside-analoga geneesmiddelen worden vaak gebruikt om virale reactivatie na HSCT te behandelen, maar hun effecten op getransplanteerde stamcellen zijn onbekend wat betreft mutageniteit en carcinogeniteit.

Concluderend, hoewel HSCT aanzienlijke vooruitgang heeft geboekt bij de behandeling van verschillende aandoeningen, is verder onderzoek nodig om transplantatieprocedures te optimaliseren, bijwerkingen te verminderen en de genomische veiligheid van getransplanteerde stamcellen te waarborgen. Technieken voor single-cell sequencing en genomische analyse worden verkend om ons begrip van het menselijke hematopoëtische systeem en de reactie op transplantatie te verbeteren.

LIST OF PUBLICATIONS

Peci F*, de Kanter JK*, Bertrums E, Rosendahl Huber A, van Leeuwen A, van Roosmalen MJ, Manders F, Verheul M, Oka R, Brandsma AM, Bierings M, Belderbos M, van Boxtel R. *Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients*. Cell Stem Cell. 2021.

Peci F*, Dekker L*, Pagliaro A, van Boxtel R, Nierkens S, Belderbos M. *The cellular composition and function of the bone marrow niche after allogeneic hematopoietic cell transplantation.* Bone Marrow Transplant. 2022.

Axel Rosendahl Huber*, Anaïs J.C. N. van Leeuwen*, **Flavia Peci**, Jurrian K. de Kanter, Eline J.M. Bertrums, and Ruben van Boxtel. *Whole-genome sequencing and mutational analysis of human cord-blood derived stem and progenitor cells.* STAR Protocols, 2022.

Sjors Middelkamp, Freek Manders*, **Flavia Peci***, Markus J. van Roosmalen, Diego Montiel González, Eline J.M. Bertrums, Inge van der Werf, Lucca L.M. Derks, Niels M. Groenen, Mark Verheul, Laurianne Trabut, Arianne M. Brandsma, Evangelia Antoniou, Dirk Reinhardt, Marc Bierings, Mirjam E. Belderbos, Ruben van Boxtel. *Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox.* bioRxiv 2023.

Maarten H. Geurts*, Shashank Gandhi*, Matteo G. Boretto*, Ninouk Akkerman, Lucca Derks, Gijs van Son, Martina Celotti, Sarina Harshuk-Shabso, **Flavia Peci**, Harry Begthel, Delilah Hendriks, Paul Schürmann, Amanda Andersson-Rolf, Susana M. Chuva de Sousa Lopes, Johan H. van Es, Ruben van Boxtel and Hans Clevers. *One-step generation of tumor models by base editor multiplexing in adult stem cell-derived organoids.* In resubmission to Nature communications

*These authors contributed equally

AUTHOR CONTRIBUTION PER CHAPTER

CHAPTER 1: General introduction and thesis outline
F.P. wrote the chapter, supervised by R.v.B.

CHAPTER 2: The cellular composition and function of the bone marrow niche after allogeneic hematopoietic cell transplantation
F.P. and L.D. wrote the manuscript, supervised by M.B.

CHAPTER 3: Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients
F.P. performed sample isolation. F.P. performed fluorescence-activated cell sorting (FACS). F.P. performed clonal expansion and supervised sequencing. J.K.d.K., M.B. and R.v.B. wrote the manuscript. J.K.d.K. performed the bioinformatic analysis. M.B. collected patient material. R.v.B. and M.B. designed and supervised the study. All authors reviewed the study.

CHAPTER 4: Whole-genome sequencing and mutational analysis of human cord-blood derived stem and progenitor cells
A.R.H. and A.J.C.N.v.L. performed the experiments. A.R.H. performed the bioinformatic analyses. F.P, A.R.H., E.J.M.B, J.K.d.K. and A.J.C.N.v.L, wrote the manuscript. F.P, A.R.H., E.J.M.B, J.K.d.K. read and revised the manuscript, all authors approved the manuscript.

CHAPTER 5: Genotoxicity and carcinogenic risks of antiviral nucleoside analog in hematopoietic stem cells
F.P., J.K., and R.B. wrote the manuscript. F.P. performed sample isolation. F.P. performed fluorescence-activated cell sorting (FACS). F.P. performed clonal expansion and supervised sequencing. J.K.d.K. performed bioinformatic analysis. R.v.B. supervised the work.

CHAPTER 6: Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox
S.M., M.E.B., and R.v.B. conceived and designed the study. S.M., F.M., and M.J.v.R. developed the PTATO computational pipeline. S.M., M.J.v.R., and F.M. performed computational analyses and designed the figures. F.P., E.J.M.B., I.v.d.W., L.L.M.D., L.T., A.M.B., and S.M. performed sample collection and single-cell isolations using flow sorting. D.M.G. performed SCAN2 variant filtering. N.M.G, M.V., L.T., and S.M. generated the AHH-1 cell lines used in this study. C.P.-M. cultured and harvested

intestinal organoids. S.M., L.T., N.M.G., and M.V. performed PTA. M.E.B., M.B., E.A., and D.R. arranged inclusion of the donors and collection of donor material. The manuscript was written by S.M., F.M., and R.v.B. with contributions from all authors.

CHAPTER 7: General Discussion

F.P. wrote the chapter, supervised by R.v.B.

List of Abbreviations

3TC	Lamivudine
ABC	Abacavir
ACV	Aciclovir
AZT	Zidovudine
BMA	Bone Marrow Adipocyte
BM	Bone Marrow
BVDU	Brivudine
CB-HSPCs	Cord-Blood Derived Hematopoietic Stem and Progenitor Cells
CGRP	Calcitonin Gene-Related Peptide
CMBCs	Cord Blood Mononuclear Blood Cells
CMV	Cytomegalovirus
CXCL12	C-X-C Motif Chemokine Ligand 12
DDC	Zalcitabine
DNA	Deoxyribonucleic Acid
DMSO	Dimethyl Sulfoxide
EMT	Emtricitabine
ETV	Entecavir
GATK	Genome Analysis Toolkit
GCV	Ganciclovir
GM-CSF	Granulocyte-Macrophage Colony-Stimulating Factor
HCT	Hematopoietic Cell Transplantation
HBV	Hepatitis B
HCV	Hepatitis C
HSC	Hematopoietic Stem Cell
HSCT	Hematopoietic Stem Cell Transplantation
hPSCs	Human Pluripotent Stem Cells
IC50	Half-Maximal Inhibitory Concentration
MZB	Mizoribine
MOV	Molnupirir
MSC	Mesenchymal Stromal Cell
PCV	Penciclovir
PBS	Phosphate-Buffered Saline
PTATO	Primary Template-Directed Amplification Analysis Toolbox
PTA	Primary Template-Directed Amplification
R	A programming language used for statistical analysis
RDV	Remdesivir
RAMP1	Receptor Activity Modifying Protein 1

RNA	Ribonucleic Acid
ROS	Reactive Oxygen Species
SBS	Single Base Substitutions
SCAN2	Somatic Mutation Rechecker and Filtering
SCF	Stem Cell Factor
SEC	Sinusoidal Endothelial Cell
SMURF	Somatic Mutation Rechecker and Filtering
SNS	Sympathetic Nervous System
TFV	Tenofovir
VEGFR2	Vascular Endothelial Growth Factor Receptor 2
WGS	Whole Genome Sequencing

ACKNOWLEDGEMENTS

Once upon a time, there was a girl, from a tiny little Italian town, passionate about biology and very curious about the future of humanity. This girl is today a woman and incredibly honored of how far she arrived and how much has achieved.

My PhD journey was the most incredible experience so far. Not only for the amazing scientific lessons learned, but especially for how much it taught myself about my strength and weaknesses. I'd happily say, if you manage to get through your PhD, you can "almost" do it all!

Of course, this journey would have not been possible if it wasn't for the patience, unwavering guidance, and expertise of my supervisors. I'm immensely grateful to **Dr. Ruben van Boxtel** and **Dr. MD Mirjam Belderbos**. Your mentorship has been instrumental in shaping my research and nurturing my intellectual growth. I am forever indebted to their profound insights, constructive feedback, and especially for the constant encouragement throughout this challenging yet rewarding endeavor.

I extend my heartfelt thanks to my promoter **Dr. Hans Clevers**, and my mentors **Dr. Claudia Janda** and **Dr. MD Jurgen Kuball** for their valuable feedback, suggestions, and rigorous examination of my work. Their diverse perspectives and expertise have significantly enhanced the quality of my research.

I started my PhD in the far 2019 and I was so excited to plan experiments and see my projects grow, when something totally unforeseeable and unexpected happened: the Corona pandemic. Corona hit the world, and the Netherlands also. It was a scary time; unable to travel to see my family and close friends, all by myself in a foreign country, I learned that you should always try to look for the silver lining. It was 2020, and the weather was exceptionally good for the Netherlands. Everything was closed but I did try to enjoy hiking in all national parks, have a taste of the local food and drinks. Also, I met a special person too who has followed me throughout this four years journey, my partner **Ryan**.

I am grateful to the **Princess Máxima Center for Pediatric Oncology** for providing the necessary resources, facilities, and vrijmibo support that enabled me to pursue my doctoral studies. The vibrant and unique academic

community and the numerous opportunities for collaboration and intellectual exchange have been truly enriching.

I'd like to raise a glass (of beer) for **Mark** and **Niels** who basically run the lab (☺). Without your jokes and professional help, I couldn't have completed my experiments and get on with my office day! Also, big congratulations to **Niels** for your new position as manager of the Máxima FACS facility, that's an amazing achievement!

Eline, you have been with me since day one and I absolutely appreciated you always being there. You brought a smile and a laugh to all the office every single day, even on the many rainy days. You have always been supportive and such a great friend. I wish you the best of luck for your future career as famous and successful pediatrician!

To **Jurrian**, your confidence and passion for science left a mark! You've an amazing mind and I'm grateful I had the chance to work together on different projects. You're next in line for defending, so best of luck with completing your PhD and for your future career plans!

Sjors, you are the best hybrid (wet/dry lab) postdoc the team could wish for. I enjoyed working together on the Fanconi Anemia project, sharing hi and lows, especially on the wet lab part! You're in incredible scientist and I'm sure you will get to your destination, so good luck for your future career as scientist and as dad of Lucy!

To **Markus**, my projects would not exist without your freshly baked bioinformatic pipelines! Thank you for providing incredible technical support and being the most silent but skilled bioinformatician in the team!

Inge, thank you for being such a supportive postdoc, always helpful and ready to answer any of my questions! I'm sure the future ahead looks bright, and I wish you all the best for your next adventures in San Diego, science-related and non!

Joske, the rugby player postdoc chess player and super auntie of our team. You rock! Thank you for being so real and down to heart, even though your IQ is probably one of the highest in the room (and perhaps the Maxima?!) Thanks for all the stimulating and open discussions, I really appreciated your free spirit.

Anais, thanks for your fashion style for Caribbean' sassiness, and you are the most sarcastic person in the room now that I left the office ☺. Best of luck for your PhD!

During my time in the office, I guess the most popular topic must have been food! My apologies to **Lucca** and **Laurianne**, for my endless jokes about being vegetarian and baking vegan cakes, I now recognize and value a more sustainable and vegetarian diet thanks to you! Also, thank you **Lucca** for the endless patience in supporting me and my R studio days and the chit chat as good old desk mates. I hope you will always remember and cherish your Italian heritage. I think you should really consider getting yourself an Italian boyfriend, they pay for dinners! I wish you both a successful career as researchers, constellated by lots of Cell, Nature and Science publications!

Vera and **Rico**, thank you for joining the van Boxtel group! You've been amazing colleagues and always up for a laugh! To **Vera**, sharing daily happiness and struggles of our PhD journeys has been great, and I wish you the very best on completing you PhD successfully, I have no doubts about it ☺

Alexander and **Diego**, I think the team really needed some exotic spin! Thank you for always sharing a good chat over coffee breaks, and for never missing a borrell. Thank you for sharing with everyone your authenticity and traditions, including amazing Cypriot and Mexican food!

Emma, the little one of the Van boxtel group, you've been with us for quite some time, and we will finish together! Always so spontaneous and enthusiastic, has been great to have you around!

Sofie, even though you join our lab very recently, I hope you will enjoy you PhD and have fun every step of it! Thank you for the coffee breaks (the good coffe downstairs) and sorry I won't be always there to prepare you dinner on your hockey training days! I can leave you some recipes if you want ☺

Francesco and **Anna**, belli giovani e bravi, mi mancherete! Grazie infinite per le good vibes, un vi faccio un grandissimo in bocca al lupo per il vostro dottorato ed il vostro futuro!

Tito, il tuo humor in dialetto veneto mi hanno svoltato le peggiori giornate!

Trisha, my Peggy Gou and **Britt**! You've got such a contagious smile, thank you so much for always lifting me up, even when I was just wining in front of the FACS machine and complaining about the Dutch weather, and always make time for a chat and party. Good luck to you both on completing your PhD and with your all your future! ☺

Special thanks also go to all the people that made my time in ML1 on the 4th floor. **Thomas**, **Niels**, **Riri**, **Chris**, **Amber**, **Esmee**, **Phylicia**, **Evelyn,** thank you for keeping the atmosphere always fun and sassy.

Big thanks to all members of the Drost group! **Irene**, **Maroussia**, **Arianna** and **Francisco**, **Kim** and **Michael (now PhD)**!

Yuehan, or better say Dr. Wang, you have been such a good friend and the best support I could ever ask for. Thank you for the endless coffee chats, dinner and brunch dates and making me feel welcome with your chinese wisdom and sassiness, absolutely loved that, and won't forget!

Kim, the real tall blonde beauty dancing queen of the Máxima, thank you for always making me laugh and having the best party in Utrecht – remember the Burlesque party- ! Good luck with the completion of you PhD and lots of hugs!

Big Big thank you to my dear friends and utrechters and amsterdammers, **Anna, Mara, Pedro, Mufaddal, Chaimaa, Mike, Camila, Owen, Carlo, Florentina** and **Kully**. The absolute very best party and travelling crew there is. You made me feel blessed and lifted every time and reminded me that there is nothing better than diversity and inclusion. Proud of every single one of you and definitely my best memories in the Netherlands lies with you! Lots of hugs

Agnieszka, my polish bestie and partner in crime, I'm so lucky we met! Endless summer and winter days together, vodka tastings, oysters and champaign to celebrate our milestones, party, sleep and repeat. Love you dearly! To happiness and money, Na Zdrowie!

Ecco il meglio alla fine, le mie besties di sempre, **Sara**, **Enrica** e **Simonetta**! Grazie per il vostro infinito sostegno, amore e lezioni di vita. Niente può battere good old amicizie di vecchia data, che sono famiglia, nemmeno la distanza. Grazie a tutta la mia **famiglia**, soprattutto a mio papi, il mio fan numero 1.

Though it is impossible to name every individual who has contributed to my success, I extend my sincere gratitude to everyone who has played a part, no matter how small, in shaping my academic journey.

Curriculum Vitae

Flavia Peci was born on July 31st, 1990, in San Benedetto del Tronto, Italy. In 2009, she obtained her scientific high school degree at Liceo Scientifico "Benedetto Rosetti" in San Benedetto del Tronto. In the same year, she started her Bachelor of Science in Medical Laboratory Science at the Marche Polytechnic University in Ancona, Italy. During her studies, she completed an internship in the Forensic Toxicology Section supervised by Prof. MD Raffaele Giorgetti at the Department of Forensic Medicine of Torrette Hospital, Ancona, Italy. In 2013, she obtained her Bachelor's degree.

In 2014, she moved to Perugia, Italy, to pursue her Master of Science in Medical Biotechnology at the University of Perugia. During her studies, she expanded her knowledge of hematology, immunology, and cancer biology, particularly gaining insight into their complex genetic backgrounds and the use of animal models to study them. For her master's thesis, she completed a project aimed to investigate and characterize a conditional knock-in mouse model for mutation A of NPM1 and the mutation ITD of FLT3 by studying the phenotype, gene expression, and protein expression. She obtained her Master's degree in 2016 and received an Erasmus scholarship to pursue a research experience outside of Italy. In February 2016, she started an internship in the Centre of Haematology and Regenerative Medicine at Karolinska Institutet (HERM), Stockholm, in the Myelodysplastic Syndrome (MDS) group led by Professor Eva Hellström-Lindberg. In Prof. MD Hellström-Lindberg's group, she worked on the study of MDS with ring sideroblast, a low-risk MDS which shares genetic signatures with Acute Myeloid Leukemia. Her project focused on the SF3B1 mutational status of ring sideroblasts engrafted in xenotransplanted NSG mice.

In January 2017, she applied for an MPhil in Medical Science at the University of Cambridge to deepen her knowledge of the role of the hematopoietic stem cell niche in hematological malignancies. During her studies, she completed an internship in the group of Professor Simón Méndez-Ferrer at the Cambridge Stem Cell Institute. The focus of her project was to characterize the splenic tumor microenvironment during the development of myeloproliferative neoplasm (MPN) and the neural regulation of stem and progenitor cells during migration and homing from the bone marrow to the spleen in physiological and pathological settings. Throughout this project, she became proficient in the use of flow cytometry, confocal microscopy, and ELISA immunoenzymatic assays. She gained independence in performing *in vitro* experiments and acquired

experience working with *in vivo* models of hematological malignancies. In 2018, she joined the Epithelial Tumor Biology group, headed by Dr. Walid Khaled, at the Department of Pharmacology, University of Cambridge. Here, she worked as a research assistant focused on protein science, biophysics, and structural biology to investigate and understand several protein-protein interactions implicated in the establishment and maintenance of triple-negative breast cancer and squamous cell lung cancer. In the summer of 2019, she began her PhD in the group of Dr. Ruben Van Boxtel at the Princess Máxima Center for Pediatric Oncology. During her PhD, she worked on multiple projects, focusing on the genomic safety of hematopoietic stem cell transplantation and antiviral treatment in pediatric cancer patients. The results of these studies are presented in her thesis. In terms of academic output, the first project she worked on resulted in a published manuscript during her second year, where she is listed as the first author. Additionally, she published a review on the effect of conditioning treatment in hematopoietic stem cell transplantation and contributed to a protocol paper. Furthermore, she was invited to present and promote her work in front of an international audience, as her abstracts were selected for multiple national and international congresses, namely the Dutch Hematology Congress (DHC), European Hematology Association (EHA), and European Bone Marrow Transplantation (EBMT). She was awarded the "Best Young Poster Abstracts at the 46th Annual Meeting of the European Society for Blood and Marrow Transplantation (EBMT 2020)" and shortlisted for the Springer Nature Basic Science Poster Award.