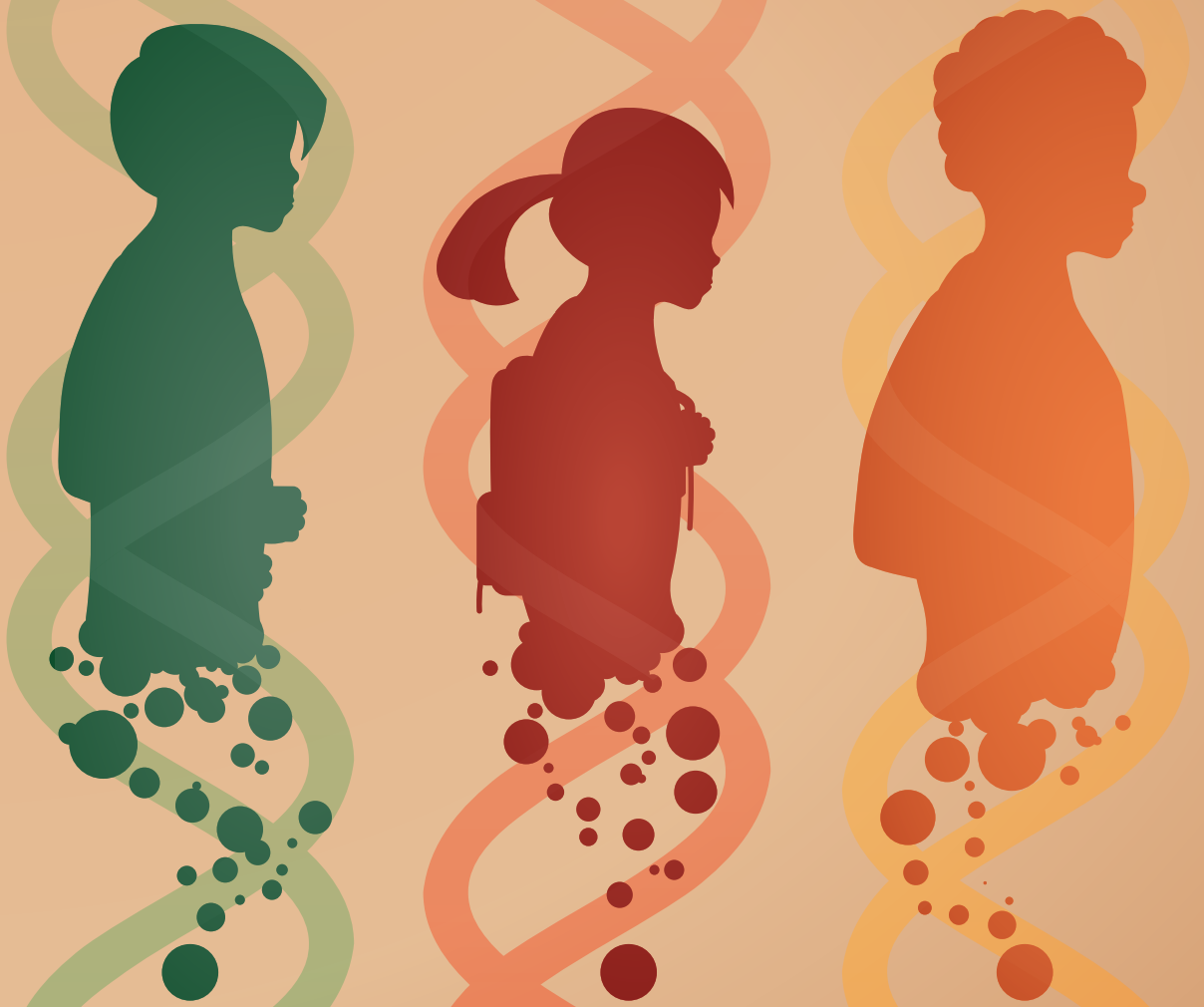
Finding the hidden patterns

single-cell omics to reduce late effects



Jurrian de Kanter

Finding the hidden patterns
single-cell omics to reduce late effects

Jurrian Kornelis de Kanter

ISBN: 978-90-393-7649-2
DOI: https://doi.org/10.33540/2219
Printed by: ProefschriftMaken.nl/ | de Bilt
Illustrations and layout: Jurrian K. de Kanter

Finding the hidden patterns
single-cell omics to reduce late effects

Het vinden van de verborgen patronen
analyse van individuele cellen om late effecten te verminderen

(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

dinsdag 9 april 2024 des middags te 2.15 uur

door

Jurrian Kornelis de Kanter

geboren op 5 oktober 1995
te Leiden

Promotor:
Prof. dr. F.C.P. Holstege

Copromotor:
Dr. R. van Boxtel

Beoordelingscommissie:
Prof. dr. ir. E.P.J.G. Cuppen
Prof. dr. L.H. Franke
Dr. J.Y. Hehir-Kwa
Prof. dr. W.L. de Laat
Prof. dr. H.J. Vormoor (voorzitter)

Table of Contents


1

Introduction: The need to study the molecular mechanisms of chemotherapy-induced late effects

Jurrian K. de Kanter[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

Childhood cancer survival in high-income countries has risen from less than 10% to more than 80% in the last six decades due to constant improvements of treatment protocols[1–4]. In the 19th century, cancer was only treated by surgery. The first significant improvements in survival were accomplished in adult patients by the addition of radiotherapy (RT) in the early 1900s, followed by the addition of the first chemotherapeutic drugs in the 1940s and 1950s[5]. As the number of pediatric cancer patients is much lower than the number of adult patients, pediatric oncologists formed large consortia that arranged clinical trials to test chemotherapies in children across hospitals and countries[4]. These consortia still exist today, and it is now standard of care in high-income countries to be enrolled in a clinical pediatric cancer trial[6]. In the first few decades, improvement of survival rates in pediatric patients was mainly achieved by adding more chemotherapeutic drugs to the treatment regimens and by intensifying the dose, especially in patients with high-risk subtypes[4]. As young children can tolerate cytotoxic compounds better than (young) adults[7], pediatric cancer patients who suffer from, among others, acute lymphoblastic leukemia (ALL), currently receive higher cumulative doses compared to adults with the same cancer type[8]. This, among other factors, results in a higher overall survival rate in pediatric compared to (young) adult cancer patients. In 2010-2014 in the Netherlands, the 5-year overall survival was 87% for ALL patients younger than 15 years old, 76% for patients 15-19 years old, and 73% for patients 20-24 years old[9]. In the United States, this was 87%, 76%, and 69% respectively[9]. However, the life-saving treatment of pediatric cancer patients comes at a cost.

Severe late effects due to cancer treatment

Chemotherapeutic compounds have various modes of action. Whereas alkylating agents directly damage the DNA, antimetabolites interfere with the replication and transcription of the DNA, and alkaloids interfere with microtubule polymerization, which is necessary for mitosis[10]. Cancer cells are generally more vulnerable to chemotherapies than most healthy cells. This vulnerability is likely caused in many cancers by their high division rate and/or their genomic instability and high mutation load[11]. The latter is specifically the case in adult cancers. This makes them susceptible to cell death induced by a further, chemotherapy-induced increase of DNA alterations via a genetic phenomenon called synthetic lethality[12].

However, healthy cells are also impacted by cancer treatment. Healthy cells can be damaged, inhibited in their proliferation, become dysfunctional, or be killed by chemotherapeutic exposure[13]. As a consequence, most cancer patients suffer from acute toxicity during treatment, such as vomiting, anemia, pain, and fatigue[14–16]. In addition, the chemotherapy-induced damage to healthy cells can lead to tissue dysfunctions that only manifest months to decades after the treatment has ended, which are collectively called late effects. These late effects include cardiovascular diseases, immunological conditions, neurological disorders, infertility, osteoporosis, and second cancers[17–19]. The latter are cancers that are genetically unrelated to the

primary cancer, and which are diagnosed during or after the treatment of the primary cancer. The biological changes in normal cells that contribute to late effects are not only caused by cancer treatment but can also be caused by a primary cancer itself. The high percentage of blasts in the blood and bone marrow of leukemia patients, for example, influences the functioning of the normal blood cells[20]. However, how a primary cancer can contribute to late effects is outside the scope of this thesis, and here the focus is on cancer treatment-induced late effects.

Across different continents, ongoing large-scale long-term follow-up cohort studies are quantifying the quality of life and the total burden of late effects of childhood cancer survivors[2,17,21]. These studies revealed that childhood cancer survivors on average develop seventeen chronic health conditions (CHCs) at the age of 50, five of which are severe, compared to nine CHCs in age-matched controls, two of which are severe[17]. Patients who receive chemo- and radiotherapy have the highest burden of CHC, while patients only treated with surgery have the lowest[19]. The class of chemotherapeutic drugs that survivors have received also influences their quality of life. For example, patients who have been treated with antimetabolites have the highest CHC burden, followed by platinum-based and alkaloid treatments, while the burden is the lowest after receiving anti-cancer antibiotics[22]. The cumulative dose of some treatments is also correlated with the incidence of specific late effects. A higher radiation dose for example is associated with a higher rate of second cancers[22,23], a higher cumulative anthracycline dose leads to high rates of cardiotoxicity[24], and a higher cumulative antimetabolite dose is associated with more late effects in the kidney compared to the treatment with any other chemotherapy[19].

The first effort to reduce late effects: decreasing the cumulative treatment dosage

The focus of clinical pediatric oncology research has undergone a gradual change over time. Initially, the only objective was to improve survival rates. However, as survival rates improved and the connection between cancer treatment and late effects became evident, clinicians also started to prioritize the long-term quality of life of survivors. The first efforts were directed toward reducing the total treatment exposure of patients with low-risk cancer subtypes. Specifically, the use of RT sharply declined at the end of the previous century, among others for the treatment of non-Hodgkin lymphoma and ALL[25]. For some cancers (e.g., Hodgkin Lymphoma) RT is still an indispensable part of treatment. However, for these cancers, the target volume that is radiated and thus the RT dose that normal tissues receive has been substantially reduced[26]. These radiation dose reductions had a clear effect. As the total number of pediatric cancer patients that received RT decreased and the median radiation dose became lower (77% with 30 Grays in the 1970s to 33% with 26 Grays in the 1990s), the incidence of late effects also decreased (from a 2.1% 15-year cumulative incidence of second cancers in the 1970s to 1.3% in the 1990s)[27]. Besides the dose of RT, the dose of chemotherapy was also reduced in subgroups of patients. The second National Wilms' Tumor Study, published in 1981, was one of the earlier

studies to show a similar efficacy of shorter chemotherapy treatment in low-risk patients, in this case six compared to fifteen months of vincristine and actinomycin D for the treatment of Wilms tumors[28]. Since then, efforts to reduce therapeutic dosages have been ongoing. Consequently, the median dose of anthracyclines, when administered, has declined over time due to its clear association with cardiotoxicity[27]. Nevertheless, when assessing all childhood cancer patients as a whole, the use of most chemotherapeutic drugs has not decreased. For example, the percentage of patients that receive anthracyclines has only increased[27]. In addition, both the frequency of administration and the dose of platinum-based drugs has strongly increased[27]. For a variety of cancer types, increasing the chemotherapy dose was essential to be able to reduce the use of RT while maintaining the same overall survival[29].

Why late effects arise: cancer treatment, aging, and DNA damage

Currently, numerous studies aim to reduce the use of chemotherapies and replace them with targeted therapies. To make these trials as effective as possible, one of the crucial steps is to expand our knowledge of the causal relationships between single chemotherapies and particular late effects. Whereas cohort study can only find the correlation between the administration of a drug and the occurrence of a late effect, molecular studies can aid in understanding the mechanism by which chemotherapies affect healthy cells. More specifically, they investigate how treatment damages or alters lipids, proteins, RNA, and DNA in healthy cells. This damage is the underlying source of all chemotherapy-induced late effects. DNA is the only one of these four damaged molecules that is not fully replaced over time. A copy of the genome, including somatically acquired DNA mutations, is passed on from each cell to its progeny. Therefore, DNA mutations accumulate in cells over time and are thus a likely source of late effects. In normal aging, mutations accumulate due to cell-intrinsic and extrinsic factors, such as replication errors, UV light, reactive oxygen species, and aldehydes[30]. In cells of cancer patients, additional DNA damage is accumulated in cells due to the genotoxic cancer treatments.

A similarity between late effects and aging does not only exist on a molecular level. Late effects also clinically resemble aging. Both aging and late effects can lead to a state that is described as frailty, i.e., diminished physiological functioning, or weakness[31]. In addition, similar diseases are associated with aging and late effects[31]. The main difference is that these conditions occur more often and at an earlier age in childhood cancer survivors[32]. As a consequence, mortality rates, caused by all leading causes of death due to aging, are higher in this group even 40 years after their diagnosis[33]. The discovery of these similarities has led to the idea that pediatric cancer survivors are subject to accelerated aging[34]. The mechanisms that drive late effects might therefore be similar to those underlying aging. The hallmarks of aging that are thought to be involved in late effects include epigenetic alterations, telomere shortening, cellular senescence, stem cell exhaustion, and genomic instability or damage[30]. There is strong evidence for the involvement of some of these mechanisms in the induction

of late effects, while only indirect evidence exists for other mechanims[35–37]. Stem cell exhaustion, for example, can lead to aging and might be induced by exposure to chemo- and radiotherapy. Direct evidence is only obtained from hematopoietic stem cell transplantation (HSCT), which has similarities to stem cell exhaustion, as a limited number of stem cells replenish the entire blood system. HSCT is associated with an increased risk of clonal hematopoiesis, (second) cancers, and cardiovascular

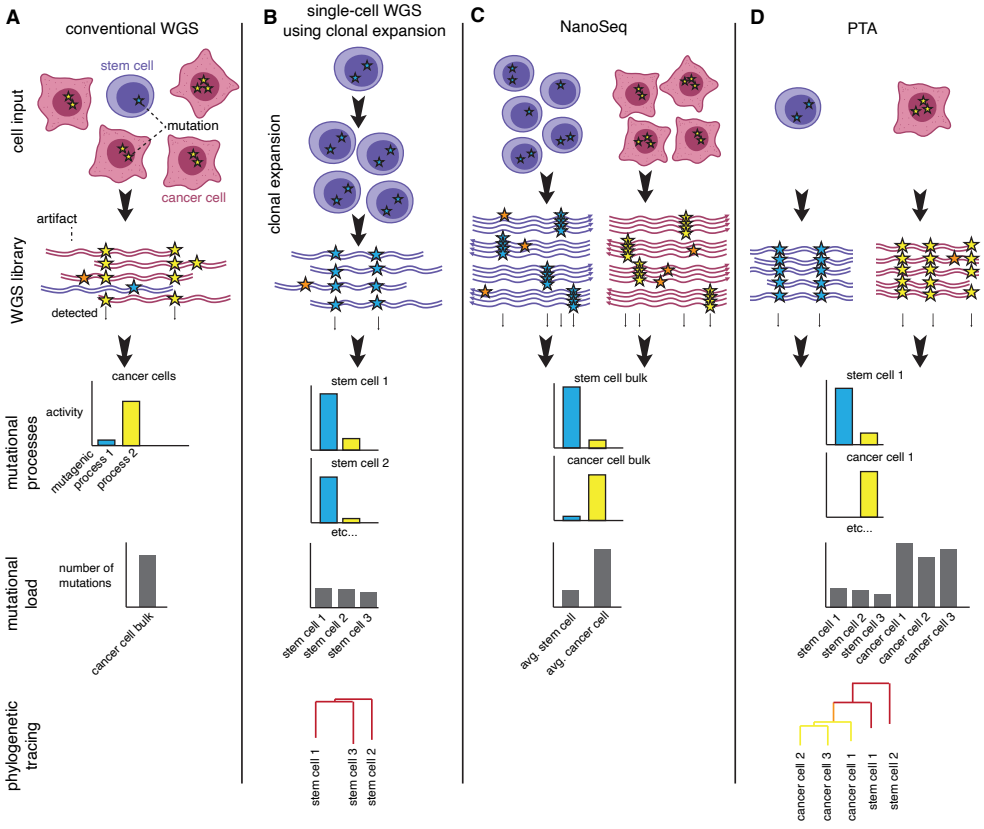


Figure 1. A schematic overview of whole genome sequencing applications
A) Conventional whole genome sequencing (WGS). The DNA of a bulk sample with a clonal population is sequenced and mutational profiles and the somatic mutation load from the major clonal population in the sample can be determined. **B)** Single stem cells can be clonally expanded *in vitro*. The DNA of the resulting clonal population can be used for conventional WGS. Using the variant allele frequency (VAF), clonal mutations from the original cell can be separated from *in vitro* acquired artifacts, resulting in per-cell mutational profiles and mutation loads. In addition, these data can be used to reconstruct a phylogenetic tree, which reveals the shared ancestry between the individual stem cells. **C)** Nano-seq can be used to detect mutations of single cells in a polyclonal population, due to the ultra-low error rate of the technique as both strands are sequenced separately. This can be used to determine the mutational processes active in any type of bulk population and to estimate the mutation load per cell. **D)** primary templated-directed amplification (PTA) uses phi29, a highly accurate DNA polymerase, in combination with exonuclease-resistant terminators to directly amplify the DNA of a single cell with high genome coverage and low artifact rates. This technique can be used to analyze the mutational processes and loads of almost any cell type and construct phylogenetic trees from these cells.

diseases, all of which are also associated with aging[38–40]. In addition, chemotherapies induce cellular senescence, at least in cancer cells[41]. Senescent cells excrete pro-inflammatory cytokines that are linked to aging-associated conditions[42]. Likely, this senescent state is also induced in some of the healthy cells in patients, but only *in vitro* and mouse experiments support this hypothesis[43–45].

Other studies, however, have directly assessed the biological aging of healthy cells in cancer survivors. For example, epigenetic age estimation is a method that estimates the biological age from a set of CpG sites of which the methylation state changes in a consistent manner over the lifetime of an individual. The epigenetic age of normal blood cells was measured in a subgroup of head and neck cancer survivors and was significantly increased 6 to 12 months after treatment[46]. In addition, in ALL survivors the level of aging-associated inflammatory cytokines is increased 5 years after the end of treatment[37]. In these patients, and in solid tumor survivors, the telomere length of lymphocytes was also significantly shorter compared to controls[36,37]. This is in line with the observation that telomerase has a decreased activity in breast cancer survivors[47]. Finally, a clear link exists between chemotherapy and accelerated genomic aging, i.e., the accumulation of additional DNA mutations. Initial assumptions that chemotherapies cause DNA aberrations in healthy cells were made based on molecular assays, such as karyograms, applied to second cancers. For example, topoisomerase inhibitors (TOPi) have long been known to induce double-strand breaks via replication stress and to be associated with 11q23 aberrations in therapy-related myeloid neoplasms (t-MN), a second cancer of the blood[48]. However, until recently, no method existed to analyze such chemotherapy-induced DNA aberrations in more detail.

How to study treatment-induced DNA mutations in cancers

Next-generation sequencing, and whole genome sequencing (WGS) in specific, have greatly enhanced the sensitivity and detail by which DNA alterations can be studied (**Fig. 1A**). For example, by sequencing and mapping the 11q23 breakpoints in t-MN, a 10bp breakpoint hotspot was found to be positioned in the KMT2A gene. The hotspot was flanked by a TOPi-induced TOP2 binding and cleavage site, strongly suggesting that the TOPi treatment directly induced these driver rearrangements[49]. Besides the more accurate assessment of structural variants, WGS also makes it possible to unbiasedly determine all small mutations, both cancer-driving and passenger. These include single base substitutions (SBS) and small insertions and deletions (indels). The genome-wide assessment of SBS and indels led to the discovery of mutational signatures, which reflect the activity of mutagenic processes (**Fig. 2A**). By taking the pyrimidine nucleotides as a reference, SBS can be grouped into six types (C>T, C>A, C>G, T>A, T>C, T>G). When also taking the preceding and following base into account, SBS can be split into 96 trinucleotide mutation types. A mutational signature is a specific pattern in the proportions of the 96 mutation types. These are recurrently detected in the genomes of multiple cancers or other samples. Using metadata such

as tissue type, the age of the patient, and behavioral and geographic data, mutational signatures can be linked to specific mutagenic processes, like aging, UV light, and smoking[50]. The same approach can be used to identify and link mutational signatures to chemotherapy drugs. This can be used in the study of late effects. For instance, in the genome of many cancer metastases and relapses that arose after exposure to platinum compounds, a recurrent SBS signature, termed "SBS31", was found (**Fig. 2B**)[51,52]. Cell line and mouse treatment experiments followed by WGS are used to validate causal relationships between mutagenic processes/agents and signatures[53,54]. A limitation of conventional WGS is that mutations can only be detected when they are shared between most cells in a sample. The clonal mutations, originally present in the initial single cell that gave rise to the clonal population, are detected in multiple reads and can be easily distinguished from artifacts and subclonal mutations and, which are both present in only one or a few reads[55,56]. WGS of tumors is effective as cancer is a clonal outgrowth by definition, with the main limitation being the percentage of normal cells infiltrating a tumor. In mouse studies, a clonal population can be obtained by inducing cancers using a carcinogen[54]. In *in vitro* studies, a clonal expansion step must be performed after treatment exposure to acquire a population that harbors clonal treatment-induced mutations. Clonal expansion is also necessary to obtain sufficient DNA from the original single cell to perform WGS (**Fig. 1B**).

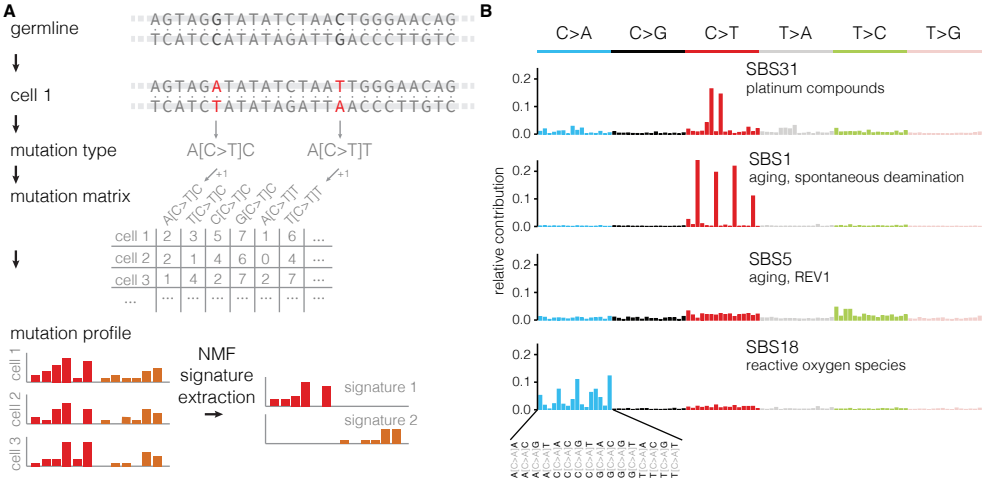


Figure 2. Construction and examples of mutational signatures

A) A schematic overview of mutational signature detection. First, the somatic mutations are identified by comparing the genome of a clonal sample or cell to a reference germline sample of the same individual and filtering out the somatic mutations from artifacts. Next, the substitution type and the preceding and following base are determined. The number of each of the 96 mutation types is counted and put in a table called a mutation matrix. The proportion of the mutation types in one sample is called the mutational profile and can be visualized in a bar plot. Finally, NMF is used to extract recurring patterns in the mutation profiles, which are called mutational signatures. **B)** The mutational signatures that are mentioned here. The proven or suspected cause of the signature is mentioned below the signature's name.

Studying how DNA damage in normal cells contributes to late effects

Application of WGS to tumor samples has yielded thorough descriptions of the number and type of mutations that are caused by chemotherapeutic agents in cancer cells and cell lines. For the study of late effects, it is important to investigate whether the same compounds are also mutagenic to normal cells *in vivo* and if so, if they result in the same type of damage, e.g., the same mutational signature. Studying the somatic mutations in normal tissues is, however, complicated by the absence of clonal expansions. As described above, a clonal population is needed to detect mutations by WGS. To circumvent this issue, stem cells from healthy tissue can be clonally expanded *in vitro*, similar to cell lines (**Fig. 1B**). This technique has been applied to stem cells of different healthy tissues such as blood, liver, and intestine, showing that the same few signatures explain all the mutations in healthy cells and that they accumulate at a constant rate over time[55–58]. Although the accumulation rate of these "clock-like" mutations is constant throughout life in all tested tissues, the exact rate of accumulation can vary between tissues. SBS1 is one of these clock-like signatures and has been linked to the spontaneous deamination of methylated cytosines at CpG dinucleotides (**Fig. 2B**)[59]. This results in T-G mismatches and C>T mutations when the mismatch remains unrepaired throughout DNA replication. This process is likely more mutagenic when a cell rapidly divides, as such a cell has less time to repair the T-G mismatch[59]. SBS5 is also a clock-like signature, which is detected in all tissues (**Fig. 2B**). A recent study suggested that the presence of SBS5 might be linked to the activity of the mutagenic translesion polymerase REV1[60]. Finally, SBS18 linearly accumulates in healthy stem cells of multiple tissues (**Fig. 2B**). It has been linked to reactive oxygen species, which can lead to the formation of 8-oxoguanines, which in turn can result in G>T mutations[61,62]. The constant accumulation of somatic mutations in healthy cells is referred to as mutational aging[63]. Mutational aging can contribute to age-related (pre-)malignant diseases, such as adult cancer and clonal hematopoiesis, and potentially also to general biological aging, for example by increasing gene-expression heterogeneity[64], although the link with biological aging is disputed[63].

A recent study clonally expanded stem cells of normal colon and liver obtained after treatment with platinum and 5-FU[65]. This revealed that, after exposure to these drugs *in vivo*, normal cells accumulate the same type of mutations as cancer cells. Even though WGS of single normal stem cells has been well established, this has only recently become possible for single differentiated cells. These cells mostly have no, or very limited, potential to expand *in vitro*, and thus not enough clonal DNA can be obtained for the minimal input required for conventional WGS. Specified culturing protocols in combination with ultra-low input WGS have recently been applied to study the mutation accumulation of healthy lymphocytes[66]. However, such an approach is more difficult to apply to highly differentiated cells like neurons and probably to cells after chemotherapy exposure, as their cycling potential is likely much decreased.

An alterative approach is using duplex sequencing, such as the NanoSeq methodology (**Fig. 1C**)[67]. In duplex sequencing, both strands of a DNA molecule are uniquely tagged and sequenced multiple times. In NanoSeq, this results in an error rate lower than 5×10^{-9} errors/bp and allows for the detection of very low-frequency mutations in polyclonal bulk samples[67]. NanoSeq can be used to detect the type of mutations that are present in a cell population and to estimate the average number of mutations in each cell. However, only 29% of the genome is covered by NanoSeq and the technique cannot determine which mutations co-occur in a cell. In 2021 Gonzalez-Pena et al. developed an alternative approach, primary template-directed amplification (PTA), in which the genome of a single cell is amplified (**Fig. 1D**)[68]. It is an adaptation of the multiple displacement amplification (MDA) protocol which uses the highly accurate phi29 DNA polymerase[69]. By adding exonuclease-resistant terminators, PTA primarily amplifies the original input DNA, which decreases amplification biases and increases the genome coverage to >90% while reducing the number of artifacts to a few hundred.

Using PTA, it is possible to analyze the genomes of single cells of virtually any cell type of a single sample and identify which mutations are shared between cells. As each cell passes its somatic mutations on to its daughter cells, shared mutations can be used to infer which cells have a common ancestor. The more mutations that two cells share, the later in life their ancestry probably separated. Inferring all clonal relationships between single cells results in an evolutionary tree and is called a phylogeny. In the context of late effects, this method can also be used to time when in the history of, for example, a t-MN, mutagenic chemotherapies were active. This can be used to gain insights into which chemotherapeutic drugs contribute to late effects and are thus prime candidates for dose reduction.

Targeted therapies to reduce late effects

In some patient groups, particularly high-risk cancer patients, reducing the treatment dose is not a viable option as it would lead to significantly poorer outcomes. In such cases, an alternative strategy to mitigate late effects is to develop compounds that possess greater selectivity in targeting cancer cells than conventional chemotherapy. Clinical trials that use targeted therapies can be more successful when strong evidence for potential targets has been gathered by molecular studies, for example, those that characterize primary cancers. Potential targets can be identified on a genotypic or phenotypic level. Genotypic targets are discovered by analyzing the DNA of primary cancers and identifying recurrently mutated genes or pathways. Phenotypic targets can be discovered by analyzing protein expression, DNA methylation, histone modifications, or gene expression in cancer. In this thesis, the focus is on therapy target identification by analyzing the transcriptome. For this, RNA sequencing (RNA-seq) is a powerful tool as it captures all mRNA transcripts in a tissue. RNA-seq data of cancer tissues can be compared to healthy tissues to find genes or pathways that have an altered expression in the cancers and that can be targeted by therapy[70].

When a target has been identified, different approaches can be taken for drug development. For example, small molecules have been developed that inhibit pathways that are overactive in cancer. These molecules differ in specificity. General kinase inhibitors, such as imatinib, are broadly applicable as they target a range of kinases that are essential for the survival of a variety of cancers (**Fig. 3A**)[71]. More specific drugs target one particular protein, which poses a vulnerability in a subset of cancers. PARP inhibitors like talazoparib, for example, can be used to treat cancers with a deficiency in the homologous recombination (HR) DNA repair pathway, via a phenomenon called synthetic lethality (**Fig. 3B**)[72]. PARP1 and PARP2 are proteins involved in the detection and repair of single-strand DNA breaks. Inhibition of PARP will lead to double strand breaks that are repaired by HR. In HR-deficient cancers this DNA damage cannot be repaired, and cancer cells die[72].

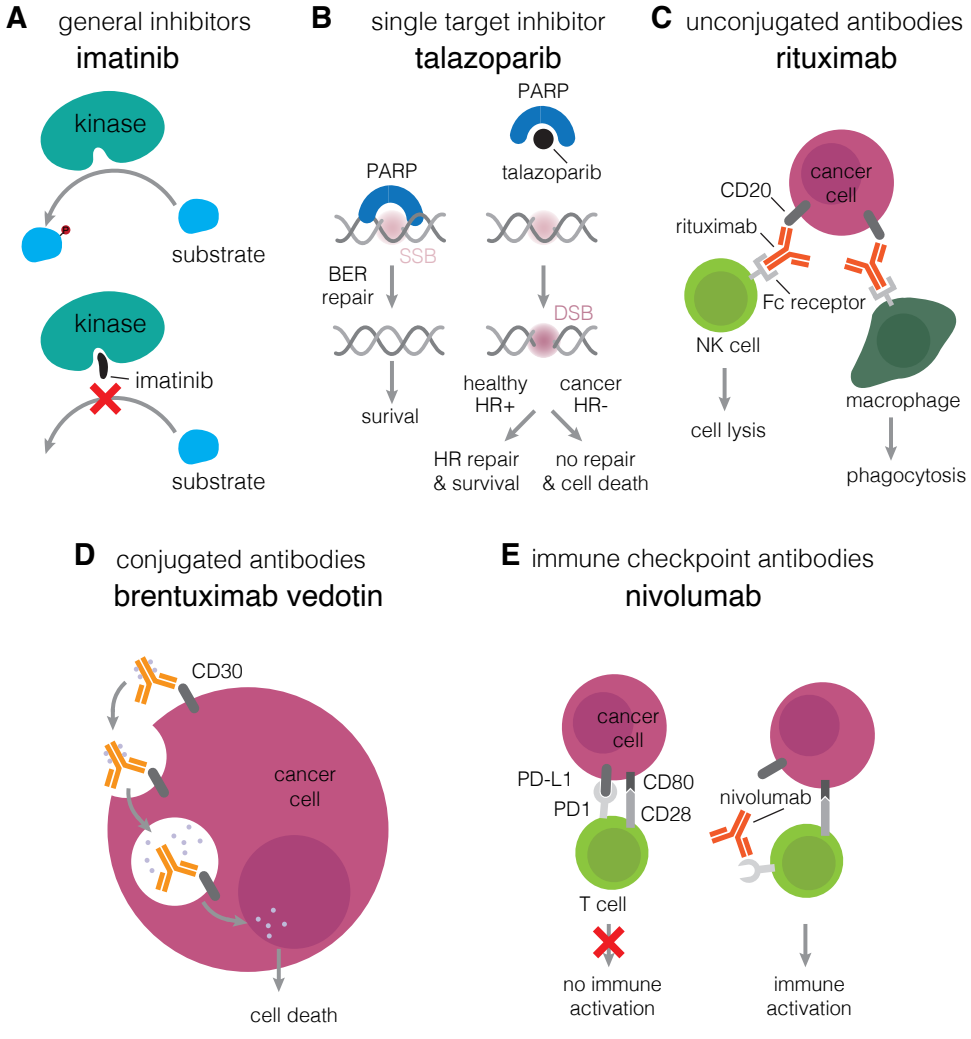


Figure 3. Working mechanisms of targeted therapies
The working mechanism of a drug of each class of targeted therapy. **A)** Imatinib, a general kinase inhibitor, inhibits the function of a group of kinases by binding their kinase domain so that they cannot phosphorylate their substrate. **B)** Talazoparib, inhibits a single protein, PARP. Normally PARP binds single-strand DNA breaks (SSB) and initiates the base excision repair (BER) pathway by which the SSB is repaired. When Talazoparib binds PARP, the BER pathway is not activated and the SSB is transformed into a double-strand DNA break (DSB). In normal cells, the DSB is repaired via the homologous repair (HR) pathway. In cancer cells that are HR deficient, e.g., due to a BRCA1/2 mutation, the DSB cannot be repaired, and the cell goes into apoptosis. **C)** Rituximab is an unconjugated antibody that binds to its target (CD20) on the surface of the cancer cell after which immune cells like NK cells and macrophages can bind the constant chain of rituximab and kill the cancer cell. **D)** Brentuximab vedotin is an antibody-drug conjugate that binds to CD30 on the surface of the cancer cell and after internalization by endocytosis ends up in lysosomes, where proteases cleave the antimitotic MMAE drug from the antibody. MMAE then blocks the polymerization of tubulin, thereby inhibiting mitosis. **E)** Nivolumab is an immune checkpoint-blocking antibody. By binding to PD1 on T cells, it inhibits the binding of PD1 to PD-L1 on the cancer cell. Normally the PD1/PD-L1 interaction inhibits activation of the T cell, but upon binding of nivolumab, the T cell can be activated, initiating an immune response against the cancer cell.

Alternatively, RNA-seq can be used to identify membrane proteins that are specifically (over-) expressed on cancer cells. These proteins can be targeted by antibodies. Rituximab, for example, binds to CD20, which is expressed on the surface of, among others, Burkitt Lymphoma (BL, **Fig. 3C**). By binding CD20 on the cancer cells, it enhances apoptosis and the killing of the cancer cells by the patient's immune system[73]. As CD20 is also expressed on healthy B-cells, this treatment can result in side effects. Antibodies targeting cancer-specific cell surface proteins can also be coupled to cytotoxic drugs, increasing the concentration of the drug in the tumor, and thereby decreasing the exposure of the rest of the body. Brentuximab vedotin (BRV), which binds CD30, a receptor that is expressed on, among others, Hodgkin Lymphoma (HL), was the second FDA-approved antibody-drug conjugate (**Fig. 3D**)[74]. Although CD30 is also expressed on a subset of healthy B cells, neural cells, and cells of the reproductive system, the addition of BRV to treatment protocols results in long-term complete remission in a subgroup of refractory or relapsed HL and better survival in a first-line regimen, while not increasing side effects[75–78].

Finally, transcriptome analysis can be used to identify immune-inhibitory ligands that are expressed by cancer cells and that bind to receptors on T cells or myeloid cells, suppressing their activity and thereby preventing the immune system from killing the cancer cells. Therapeutic antibodies have been developed that interfere with the binding of such ligand-receptor pairs, thereby increasing the immune activation against the cancer cells. Nivolumab for example interferes with PD-L1 that is expressed on some cancer cells (**Fig. 3E**)[79]. PD-L1 binds to PD-1 on T cells, suppressing T cell activity. Conventional RNA-seq is, however, limited to the analysis of a bulk population, which in most cancers consists of a mixture of malignant and normal cells. This makes it impossible when analyzing a bulk sample to confirm that both the receptor (e.g., PD-1) and the corresponding ligand (e.g., PD-L1) are expressed on the immune cells and malignant cells, respectively.

Single-cell RNA sequencing (scRNA-seq) overcomes this problem. It captures and analyzes the transcriptome of each individual cell. This way, the cancer cells and the microenvironment can be analyzed simultaneously. This allows for the accurate detection of cancer cell-immune cell interactions, together with any other category of potential treatment target[80]. Another advantage of scRNA-seq is that it allows for the detection of rare populations of cells. This can be beneficial when analyzing cancers with low fractions of malignant cells, like HL. In this cancer, most of the tumor consists of immune cells and only 0.1-5% consists of malignant cells[81]. Using conventional RNA-seq, this rare population cannot be accurately studied when processing the bulk tumor sample. However, using scRNA-seq it is possible to capture this minor fraction of malignant cells concurrently with the immune cells.

Thesis scope and outline

As described in this chapter, current treatment regimens for childhood cancer result in a high burden of late effects. The work described in this thesis aims to expand the molecular knowledge needed for the design of clinical research on late effects reduction. Two approaches are taken. First, the mutagenicity of chemotherapeutic compounds was investigated in healthy cells and second cancers. Drugs that are highly mutagenic to healthy cells would be prime candidates for dose reduction or replacement. Second, candidate genes for targeted therapy development were identified in primary cancers. These could potentially be used to develop novel treatments that induce fewer late effects. Both topics have been the focus of extensive research for the past few decades. The recently developed single-cell genomic and transcriptomic methods used in this work have opened up new avenues to investigate these topics in unprecedented detail.

In most treatment protocols, chemo- and radiotherapy are the first line of therapy. These lead to late effects like secondary cancers, among which are therapy-related myeloid neoplasms (t-MN). In **chapter 2** we applied whole genome sequencing (WGS) to t-MN and normal, clonally expanded hematopoietic stem and progenitor cells (HSPCs). We found that chemotherapies increase the number of mutations in t-MN and normal HSPCs. Only a few chemotherapies directly induced mutations in healthy cells, while most drugs indirectly caused an increase in the number of clock-like mutations. Phylogenetic lineage tracing indicated that most t-MN originated after the start of treatment and became dominant during or after treatment. In **chapter 3** we applied PTA to sequence both healthy HSPCs and single t-MN blasts to study the selective pressures that treatments induce on (pre-)leukemic clones. This revealed that platinum-induced inhibition of cell division is the rate-limiting step of t-MN expansion in platinum-treated patients. In addition, we showed that this inhibition is likely TP53 dependent and that fully TP53-deficient t-MN can divide more efficiently under platinum treatment, shortening their latency.
Hematopoietic stem cell transplantation (HSCT) can be applied as the second line of treatment for blood cancers and is associated with clonal hematopoiesis, t-MN, and

cardiovascular late effects[38–40]. In **chapter 4**, we confirmed the general genomic safety of HSCT by sequencing normal HSPCs before and after the procedure and finding a similar mutational load in all donor HSPCs and most recipient HSPCs. However, we found that the treatment of HSCT-related viral infections with the antiviral nucleoside analog ganciclovir can lead to the accumulation of additional mutations in HSPCs. These mutations were also found in cancers where they led to cancer-driving events. In **chapter 5**, we applied an *in vitro* treatment method followed by WGS to screen the mutagenicity of a compendium of nucleoside analogs in healthy, uninfected human cells. While most nucleoside analogs were not mutagenic, five did induce mutations, although none as many as ganciclovir. In conclusion, we have gained evidence that specific chemotherapies and antiviral drugs contribute to the development of second cancers. These drugs could therefore be potential candidates for drug replacement or decreased usage.

When dose reduction is not an option, the substitution of current chemotherapies by targeted treatment can be used to reduce late effects. Such an approach was successfully applied to patients with HL, which is associated with some of the most severe late effects of any cancer[17]. However, these targeted therapies are still combined with high-dose chemotherapy[77–79,82]. Here, we characterized pediatric HL by single-cell RNA-sequencing and identified potential therapeutic targets to further improve the (combined) efficacy of targeted therapies. One challenge in the analysis of single-cell RNA sequencing is the classification of cell types based on the transcriptional profiles of individual cells. In **chapter 6,** we present CHETAH, a robust, automated cell-type classification algorithm for scRNA-seq data that is able to distinguish cancer cells from normal cells in the tumor microenvironment (TME). In **chapter 7**, we applied scRNA-seq on pediatric HL and used CHETAH to identify the cell types in the data. We found that NK cells and subtypes of T cells are likely suppressed by HRS cells via different inhibitory receptors, but that the interaction strengths vary per patient. We used RNAscope to validate these interactions and show that there is high inter- and intra-patient variability in the strength of these interactions.

In **chapter 8**, the insights that can be extracted from the combined work described in this thesis are discussed. In addition, the potential long-term clinical impact of the findings from this thesis is discussed and the steps needed to reach that goal. Finally, the effectiveness of the research is reviewed and future research directions are presented.

Together, the work described in this thesis contributes to our understanding of the origin of late effects and adds to the fundamental knowledge that is needed todevelop treatment regimens that are less toxic to normal tissues and therefore lead to fewer late effects.

References

1. Gatta, G. et al. Childhood cancer survival in Europe 1999-2007: Results of EUROCARE-5-a population-based study. Lancet Oncol 15, 35–47 (2014).
2. Youlden, D. R. et al. Childhood cancer survival and avoided deaths in Australia, 1983–2016. Paediatr Perinat Epidemiol 37, 81–91 (2023).
3. Zhao, J. et al. Racial/ethnic disparities in childhood cancer survival in the United States. Cancer Epidemiology Biomarkers and Prevention 30, 2010–2017 (2021).
4. O'Leary, M., Krailo, M., Anderson, J. R. & Reaman, G. H. Progress in Childhood Cancer: 50 Years of Research Collaboration, a Report From the Children's Oncology Group. Semin Oncol 35, 484–493 (2008).
5. DeVita, V. T. & Chu, E. A History of Cancer Chemotherapy. Cancer Res 68, 8643–8653 (2008).
6. Unguru, Y. The successful integration of research and care: How pediatric oncology became the subspecialty in which research defines the standard of care. Pediatr Blood Cancer 56, 1019–1025 (2011).
7. Neaga, A. et al. Why Do Children with Acute Lymphoblastic Leukemia Fare Better Than Adults? Cancers (Basel) 13, 3886 (2021).
8. Ram, R. et al. Adolescents and young adults with acute lymphoblastic leukemia have a better outcome when treated with pediatric-inspired regimens: Systematic review and meta-analysis. Am J Hematol 87, 472–478 (2012).
9. Ssenyonga, N. et al. Worldwide trends in population-based survival for children, adolescents, and young adults diagnosed with leukaemia, by subtype, during 2000–14 (CONCORD-3): analysis of individual data from 258 cancer registries in 61 countries. Lancet Child Adolesc Health 6, 409–431 (2022).
10. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet 15, 585–598 (2014).
11. Caley, A. & Jones, R. The principles of cancer treatment by chemotherapy. Surgery (Oxford) 30, 186–190 (2012).
12. Fox, E. J. & Loeb, L. A. Lethal Mutagenesis: Targeting the Mutator Phenotype in Cancer. Semin Cancer Biol 20, 353–359 (2010).
13. Wang, S., Prizment, A., Thyagarajan, B. & Blaes, A. Cancer Treatment-Induced Accelerated Aging in Cancer Survivors: Biology and Assessment. Cancers (Basel) 13, 427 (2021).
14. Sherani, F., Boston, C. & Mba, N. Latest Update on Prevention of Acute Chemotherapy-Induced Nausea and Vomiting in Pediatric Cancer Patients. Curr Oncol Rep 21, 89 (2019).
15. Alessi, I. et al. Short and Long-Term Toxicity in Pediatric Cancer Treatment: Central Nervous System Damage. Cancers (Basel) 14, 1540 (2022).
16. Schmiegelow, K. et al. Consensus definitions of 14 severe acute toxic effects for childhood lymphoblastic leukaemia treatment: a Delphi consensus. Lancet Oncol 17, e231–e239 (2016).
17. Bhakta, N. et al. The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE). The Lancet 390, 2569–2582 (2017).
18. Erdmann, F. et al. Childhood cancer: Survival, treatment modalities, late effects and improvements over time. Cancer Epidemiol 71, (2021).
19. Chang, W. H. et al. Late effects of cancer in children, teenagers and young adults: Population-based study on the burden of 183 conditions, in-patient and critical care admissions and years of life lost. The Lancet Regional Health - Europe 12, 100248 (2022).
20. Batsivari, A., Grey, W. & Bonnet, D. Understanding of the crosstalk between normal residual hematopoietic stem cells and the leukemic niche in acute myeloid leukemia. Exp Hematol 95, 23–30 (2021).
21. Mulder, R. L. et al. Fertility preservation for female patients with childhood, adolescent, and young adult cancer: recommendations from the PanCareLIFE Consortium and the International Late Effects of Childhood Cancer Guideline Harmonization Group. Lancet Oncol 22, e45–e56 (2021).
22. Mittal, A. et al. Late effects in pediatric Hodgkin lymphoma survivors after uniform treatment with ABVD with or without radiotherapy. Pediatr Blood Cancer 68, 1–12 (2021).
23. Lee, J. S. et al. Increased risk of second malignant neoplasms in adolescents and young adults with cancer. Cancer 122, 116–123 (2016).
24. Hudson, M. M. et al. Noninvasive evaluation of late anthracycline cardiac toxicity in childhood cancer survivors. Journal of Clinical Oncology 25, 3635–3643 (2007).
25. Jairam, V., Roberts, K. B. & Yu, J. B. Historical trends in the use of radiation therapy for pediatric cancers: 1973-2008. Int J Radiat Oncol Biol Phys 85, (2013).
26. Specht, L. Radiotherapy for Hodgkin Lymphoma. The Cancer Journal 24, 237–243 (2018).
27. Turcotte, L. M. et al. Temporal Trends in Treatment and Subsequent Neoplasm Risk Among 5-Year Survivors of Childhood Cancer, 1970-2015. JAMA 317, 814 (2017).
28. D'Angio, G. J. et al. The treatment of Wilms' Tumor: Results of the second national Wilms' Tumor study.

Cancer 47, 2302–2311 (1981).
29. Hudson, M. M. et al. Lessons from the past: Opportunities to improve childhood cancer survivor care through outcomes investigations of historical therapeutic approaches for pediatric hematological malignancies. Pediatr Blood Cancer 58, 334–343 (2012).
30. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. Cell 153, 1194–1217 (2013).
31. Smitherman, A. B. et al. Accelerated aging among childhood, adolescent, and young adult cancer survivors is evidenced by increased expression of p16INK4a and frailty. Cancer 126, 4975–4983 (2020).
32. Carroll, J. E., Bower, J. E. & Ganz, P. A. Cancer-related accelerated ageing and biobehavioural modifiers: a framework for research and clinical care. Nat Rev Clin Oncol 19, 173–187 (2022).
33. Dixon, S. B. et al. Specific causes of excess late mortality and association with modifiable risk factors among survivors of childhood cancer: a report from the Childhood Cancer Survivor Study cohort. The Lancet 401, 1447–1457 (2023).
34. Henderson, T. O., Ness, K. K. & Cohen, H. J. Accelerated Aging among Cancer Survivors: From Pediatrics to Geriatrics. American Society of Clinical Oncology Educational Book e423–e430 (2014) doi:10.14694/EdBook_AM.2014.34.e423.
35. Cupit-Link, M. C. et al. Biology of premature ageing in survivors of cancer. ESMO Open 2, e000250 (2017).
36. Franco, S. et al. Telomere dynamics in childhood leukemia and solid tumors: A follow-up study. Leukemia 17, 401–410 (2003).
37. Ariffin, H. et al. Young adult survivors of childhood acute lymphoblastic leukemia show evidence of chronic inflammation and cellular aging. Cancer 123, 4207–4214 (2017).
38. Danylesko, I. & Shimoni, A. Second Malignancies after Hematopoietic Stem Cell Transplantation. Curr Treat Options Oncol 19, 9 (2018).
39. López-Fernández, T., Vadillo, I. S., de la Guía, A. L. & Barbier, K. H. Cardiovascular Issues in Hematopoietic Stem Cell Transplantation (HSCT). Curr Treat Options Oncol 22, 51 (2021).
40. Frick, M. et al. Role of Donor Clonal Hematopoiesis in Allogeneic Hematopoietic Stem-Cell Transplantation. Journal of Clinical Oncology 37, 375–385 (2019).
41. Duy, C. et al. Chemotherapy Induces Senescence-Like Resilient Cells Capable of Initiating AML Recurrence. Cancer Discov 11, 1542–1561 (2021).
42. Guida, J. L. et al. Measuring Aging and Identifying Aging Phenotypes in Cancer Survivors. JNCI: Journal of the National Cancer Institute 111, 1245–1254 (2019).
43. Di, X. et al. A chemotherapy-associated senescence bystander effect in breast cancer cells. Cancer Biol Ther 7, 864–872 (2008).
44. Sapega, O. et al. Distinct phenotypes and 'bystander' effects of senescent tumour cells induced by docetaxel or immunomodulatory cytokines. Int J Oncol 53, 1997–2009 (2018).
45. Demaria, M. et al. Cellular Senescence Promotes Adverse Effects of Chemotherapy and Cancer Relapse. Cancer Discov 7, 165–176 (2017).
46. Xiao, C. et al. Association of Epigenetic Age Acceleration With Risk Factors, Survival, and Quality of Life in Patients With Head and Neck Cancer. Int J Radiation Oncol Biol Phys 111, 2021 (2021).
47. Scuric, Z. et al. Biomarkers of aging associated with past treatments in breast cancer survivors. NPJ Breast Cancer 3, (2017).
48. Prieto, F. et al. 11q23 abnormalities in children with acute nonlymphocytic leukemia (M4–M5): Association with previous chemotherapy. Cancer Genet Cytogenet 45, 1–11 (1990).
49. Mirault, M.-E., Boucher, P. & Tremblay, A. Nucleotide-Resolution Mapping of Topoisomerase-Mediated and Apoptotic DNA Strand Scissions at or near an MLL Translocation Hotspot. The American Journal of Human Genetics 79, 779–791 (2006).
50. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
51. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. Nature 578, 94–101 (2020).
52. Schwartz, J. R. et al. The acquisition of molecular drivers in pediatric therapy-related myeloid neoplasms. Nat Commun 12, 985 (2021).
53. Kucab, J. E. et al. A Compendium of Mutational Signatures of Environmental Agents. Cell 177, 821-836.e16 (2019).
54. Riva, L. et al. The mutational signature profile of known and suspected human carcinogens in mice. Nat Genet 52, 1189–1197 (2020).
55. Osorio, F. G. et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep 25, 2308-2316.e4 (2018).
56. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. Nature 561, 473–478 (2018).
57. Jager, M. et al. Measuring mutation accumulation in single human adult stem cells by whole-genome

sequencing of organoid cultures. Nat Protoc 13, 59–78 (2018).
58. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 574, 532–537 (2019).
59. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. Nat Genet 47, 1402–1407 (2015).
60. Petljak, M. et al. Mechanisms of APOBEC3 mutagenesis in human cancer cells. Nature 607, 799–807 (2022).
61. Jin, S.-G., Meng, Y., Johnson, J., Szabó, P. E. & Pfeifer, G. P. Concordance of hydrogen peroxide–induced 8-oxo-guanine patterns with two cancer mutation signatures of upper GI tract tumors. Sci Adv 8, eabn3815 (2023).
62. van den Boogaard, M. L. et al. Defects in 8-oxo-guanine repair pathway cause high frequency of C > A substitutions in neuroblastoma. PNAS 118, e2007898118 (2021).
63. Robinson, P. S. et al. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. Nat Genet 53, 1434–1442 (2021).
64. Manders, F., van Boxtel, R. & Middelkamp, S. The Dynamics of Somatic Mutagenesis During Life in Humans. Frontiers in Aging 2, (2021).
65. Kuijk, E., Kranenburg, O., Cuppen, E. & Van Hoeck, A. Common anti-cancer therapies induce somatic mutations in stem cells of healthy tissue. Nat Commun 13, 5915 (2022).
66. Machado, H. E. et al. Diverse mutational landscapes in human lymphocytes. Nature 608, 724–732 (2022).
67. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. Nature 593, 405–410 (2021).
68. Gonzalez-Pena, V. et al. Accurate genomic variant detection in single cells with primary template-directed amplification. Proc Natl Acad Sci U S A 118, 1–12 (2021).
69. Spits, C. et al. Whole-genome multiple displacement amplification from single cells. Nat Protoc 1, 1965–1970 (2006).
70. Hong, M. et al. RNA sequencing: new technologies and applications in cancer research. J Hematol Oncol 13, 166 (2020).
71. Iqbal, N. & Iqbal, N. Imatinib: A Breakthrough of Targeted Therapy in Cancer. Chemother Res Pract 2014, 1–9 (2014).
72. Lord, C. J. & Ashworth, A. PARP inhibitors: Synthetic lethality in the clinic. Science (1979) 355, 1152–1158 (2017).
73. Weiner, G. J. Rituximab: Mechanism of Action. Semin Hematol 47, 115–123 (2010).
74. Thomas, A., Teicher, B. A. & Hassan, R. Antibody–drug conjugates for cancer therapy. Lancet Oncol 17, e254–e262 (2016).
75. Younes, A. et al. Results of a pivotal phase II study of brentuximab vedotin for patients with relapsed or refractory Hodgkin's lymphoma. Journal of Clinical Oncology 30, 2183–2189 (2012).
76. Chen, R. et al. Five-year survival and durability results of brentuximab vedotin in patients with relapsed or refractory Hodgkin lymphoma. Blood 128, 1562–1566 (2016).
77. Castellino, S. M. et al. Brentuximab Vedotin with Chemotherapy in Pediatric High-Risk Hodgkin's Lymphoma. New England Journal of Medicine 387, 1649–1660 (2022).
78. Ansell, S. M. et al. Overall Survival with Brentuximab Vedotin in Stage III or IV Hodgkin's Lymphoma. New England Journal of Medicine 387, 310–320 (2022).
79. Herrera, A. F. et al. Interim results of brentuximab vedotin in combination with nivolumab in patients with relapsed or refractory Hodgkin lymphoma. Blood 131, 1183–1194 (2018).
80. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. Nat Rev Genet 22, 71–88 (2021).
81. Weniger, M. A. & Küppers, R. Molecular biology of Hodgkin lymphoma. Leukemia 35, 968–981 (2021).
82. Herrera, A. F. et al. Brentuximab vedotin plus nivolumab after autologous haematopoietic stem-cell transplantation for adult patients with high-risk classic Hodgkin lymphoma: a multicentre, phase 2 trial. Lancet Haematol 10, e14–e23 (2023).


2

Elevated mutational age in blood of children treated for cancer contributes to therapy-related myeloid neoplasms

Eline J.M. Bertrums[1,2,3,*], Axel K.M. Rosendahl Huber[1,2,*], **Jurrian K. de Kanter**[1,2,*], Arianne M. Brandsma[1,2], Anaïs J.C.N. van Leeuwen[1,2], Mark Verheul[1,2], Marry M. van den Heuvel-Eibrink[1,4], Rurika Oka[1,2], Markus J. van Roosmalen[1,2], Hester A. de Groot-Kruseman[1], C. Michel Zwaan[3,1], Bianca F. Goemans[1] , 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
[3] Department of Pediatric Oncology, Erasmus Medical Center – Sophia Children's Hospital, Rotterdam, the Netherlands
[4] Utrecht University, Utrecht, the Netherlands
* These authors contributed equally

Cancer Discovery, 2022 Aug 5;12(8):1860-1872.
doi: 10.1158/2159-8290.CD-22-0120.

Abstract

Childhood cancer survivors are confronted with various chronic health conditions like therapy-related malignancies. However, it is unclear how exposure to chemotherapy contributes to the mutation burden and clonal composition of healthy tissues early in life. Here, we studied mutation accumulation in hematopoietic stem and progenitor cells (HSPCs) before and after cancer treatment of 24 children. Of these, 19 developed therapy-related myeloid neoplasms (t-MNs). Posttreatment HSPCs had an average mutation burden increase comparable to what treatment-naive cells accumulate during 16 years of life, with excesses up to 80 years. In most children, these additional mutations were induced by clock-like processes, which are also active during healthy aging. Other patients harbored mutations that could be directly attributed to treatments like platinum-based drugs and thiopurines. Using phylogenetic inference, we demonstrate that most t-MN in children originate after the start of treatment and that leukemic clones become dominant during or directly after chemotherapy exposure.

Introduction

Chemotherapy is the major treatment modality for cancer and has led to curative treatment of an increasingly large number of patients[1]. Most chemotherapeutic drugs act by fatally damaging or blocking DNA replication in malignant cells[2]. However, chemotherapy can also be highly mutagenic to malignant cells that survive the cytotoxic effects[3]. Indeed, whole-genome sequencing (WGS) analyses of more than 3,500 cancer metastases revealed several mutational signatures, which are a direct consequence of exposure to specific chemotherapeutic drugs, such as platinum-based compounds and 5-fluorouracil (5-FU)[4,5]. In addition, experimental strategies have defined and confirmed mutational signatures induced by chemotherapy *in vitro*, such as temozolomide, platinum-based compounds, cyclophosphamide, 5-FU and 6-mercaptopurine (6-MP)[6-8].

During chemotherapeutic treatment, the toxic effects on normal tissues are often dose limiting, the hematopoietic system being especially vulnerable[9]. Besides these acute toxic effects, cancer survivors are confronted with a variety of chronic health conditions later in life as a result of chemotherapy, such as cardiac problems, infertility, and secondary malignancies[10-12]. Especially childhood cancer survivors suffer from these long-term adverse effects, which collectively resemble accelerated aging[12]. Indeed, while the long-term survival rate of children treated for cancer approaches 80%, their bodies are still in development during treatment and survivors can develop adverse effects even decades after their initial diagnosis[12]. Chemotherapy-induced mutagenesis and clonal expansions in healthy tissues may be responsible for inducing some of these long-term adverse effects, in particular secondary malignancies. Thus far, the impact of chemotherapy exposure in normal blood has been inferred from mutational landscapes of therapy-related myeloid neoplasms (t-MN)[5,13] as well as of

cases of clonal hematopoiesis (CH) in exposed patients[14]. t-MN comprises two disease types, namely therapy-related acute myeloid leukemia (t-AML) and therapy-related myelodysplastic syndrome (t-MDS)[15,16]. In t-MN genomes, high numbers of clonal platinum- and thiopurine-induced mutations could be observed, which indicated that in these cases clonal expansion of the hematopoietic cell founding the leukemia started after the initiation of exposure[5]. Indeed, cancer therapy preferentially selects for cells that harbor mutations in DNA damage response genes, such as *TP53*, *CHEK2* and *PPM1D*, ultimately resulting in CH and t-MN[5]. In adults, preleukemic CH can often already be observed at the time of the primary cancer diagnosis and before exposure to treatment[5,17]. However, a recent report on three pediatric neuroblastoma patients showed that most mutations present in the CH clone in these children could be linked to platinum-associated signatures[16]. These previous studies focused on studying mutational landscapes of preleukemic clonal expansions and/or t-MN. However, the mutational consequences of chemotherapy exposure in normal HSPCs and how this relates to mutations observed in t-MN is unknown. In addition, the origin of t-MN with respect to the timing of chemotherapy exposure in children seems different from adults yet remains understudied.

Here, we characterized the mutational consequences of chemotherapy in normal HSPCs of pediatric cancer patients before and after receiving treatment. We included patients who developed t-MN to determine how mutagenesis and clonal evolution in healthy blood contribute to the development of secondary malignancies. We found that the mutation burden of normal HSPCs was increased after chemotherapy. Remarkably, chemotherapy-associated mutagenesis in most patients was caused by processes resembling those active during healthy aging. Only few compounds, such as platinum-based drugs and thiopurines, had a direct mutagenic impact. In contrast to the effect of thiopurines, our data suggest that the effect of platinum-based drugs is independent of cell division. By combining mutational signature analysis and phylogenetic inference, we demonstrate that both induction of driving events and subsequent selection occur during chemotherapeutic exposure, which can ultimately lead to t-MN.

Results

Cataloguing somatic mutations in chemotherapy-exposed HSPCs of children

To assess the mutational consequences of chemotherapy exposure in normal tissues, we determined the somatic mutations present in HSPCs of children before and after receiving cancer treatment. We focused on the hematopoietic system, since it is highly sensitive to chemotherapy exposure[9] and because the lifelong mutation accumulation in healthy HSPCs has been determined[18-20]. In total, we assessed 24 patients, who underwent treatment for different pediatric cancer types (**Fig. 1A**). Of these, 19 patients developed t-MN (**Fig. 1B**), of which 18 were t-AML and one was t-MDS (UPN012). A detailed list of diagnosis, age, and treatment information for all patients is provided in **Table S1**. The latency between the start of chemotherapy and t-MN onset among patients ranged from 1.1 to 10 years (**Fig. 1B**).

Depending on the available patient material, we performed WGS on individual HSPCs at the time of the primary diagnosis (DX1), after complete remission of this first cancer (FU; posttreatment), and at the time of t-MN diagnosis (DX2) (**Fig. 1B,C**). In addition, we sequenced the t-MN genomes. Of note, the t-MN of patient UPN021 was excluded because of poor sequencing quality (Methods). (**Fig. 1A, Table S2**). We sorted single HSPCs using flow cytometry (Methods), clonally expanded these cells *in vitro* to obtain sufficient DNA for WGS (**Fig. 1C, Fig. S1**) and catalogued all clonal somatic mutations per HSPC (Methods, **Fig. S2A**).

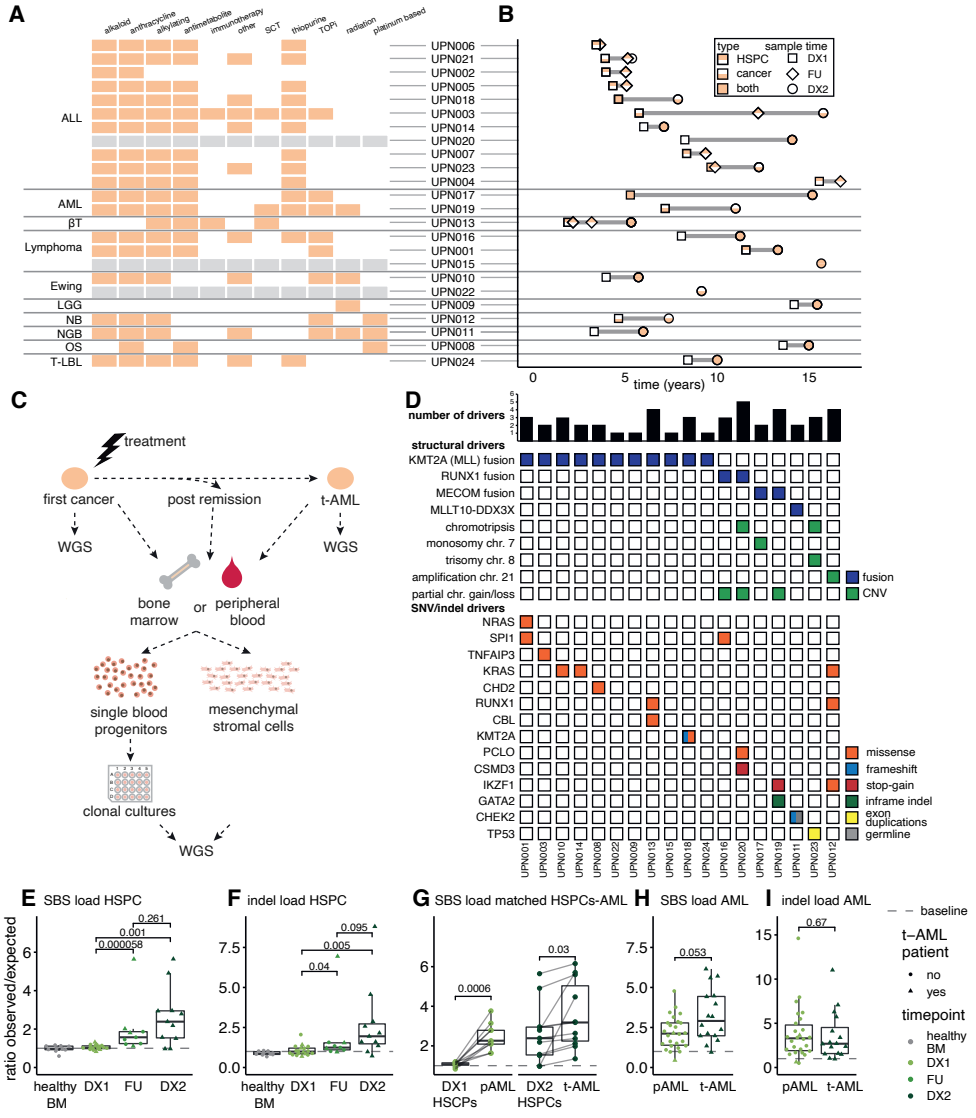


Figure 1. HSPCs and t-MN blasts in a chemotherapy-treated patient cohort display an increased mutation load.

Legend on the next page

A) A table depicting the different treatment categories that patients from each tumor type received. Rows per patient match with timelines in B. ALL, acute lymphoblastic leukemia; βT, beta-thalassemia; LGG, low grade glioma; NB, neuroblastoma; NGB, neuroganglioblastoma; OS, osteosarcoma; SCT, stem cell transplantation; T-LBL, T-cell lymphoblastic lymphoma; TOPi, topoisomerase inhibitor. **B)** The per-patient timelines of sample collection and the type of material that was sequenced. **C)** A schematic overview of the experimental setup of this study. In short, biopsies at time of the primary cancer, follow-up (after remission) and t-MN were collected. Blasts were enriched by FACS, mesenchymal stromal cells were expanded *in vitro* and both were sequenced in bulk. FACS was also used to sort single HSPCs into 96-well plates, which were then clonally expanded to obtain sufficient DNA for WGS, after which the mutation catalogues of the original HSPCs could be obtained. **D)** The clonal driver events were identified in the t-MN samples. The bar plots on top indicate the number of driver events identified in each sample. **E)** The mutation load of single base substitutions (SBS) per time point per HSPC, normalized to the HSPC baseline consisting of healthy bone marrow (BM) HSPCS. The HSPCs were averaged per patient per timepoint. Two-sided Wilcox-test, FDR-corrected. Here, and in all other figures, the boxplots depict the median (center line), 25th and 75th percentiles (box), and the largest values no more than 1.5* the interquartile range (whiskers). **F)** similar to E but for indels in HSPCs. **G)** The mutation load of single base substitutions of primary and therapy-related AML blasts and the mean value of matched HSPCs from the same timepoint per patient. Connecting lines represented matched AML and HSPCs. Two-sided paired t test, FDR-corrected. **H)** Similar to E but for t-AML blasts. **I)** Similar to F but for t-AML blasts.

In total, we assessed 135 HSPC clones (28 before and 107 after treatment) and identified 46,831 single base substitutions, 2,658 small insertions and deletions (indels) and 346 double base substitutions. Most of the t-MN cases were driven by gene fusions (16 out of 18 cases; **Fig. 1D**). Of these, 11 patients (69%) harbored a MLL fusion (KMT2A rearrangement), two patients (13%) a RUNX1 fusion and two patients (13%) a MECOM fusion (**Fig. 1D**, **Fig. S2B**). Five MLL breakpoints were present in the 11bp topoisomerase II inhibitor (TOP2i)-related hotspot positioned within the MLL breakpoint cluster region, and all of these patients received TOP2i[21,22] (**Fig. S2C**). In four patients the MLL fusion was the sole t-MN driver. The frequency of MLL fusions in these pediatric t-MN patients is considerably higher than previously reported in primary pediatric and infant AML (13% and 38%; respectively)[23] and in adult t-AML (23%)[24]. Nonetheless, we did identify genetic similarities between primary pediatric AML (pAML) and t-AML, such as a higher prevalence of RAS mutations in MLL-rearranged (MLLr) AML (**Fig. 1D**). Finally, despite the use of alkylating agents in many of our patients and its previously described association with monosomy 5(q) and 7(q)[13,24], only three t-MNs (17%) presented with this aberration. The lack of t-MN cases with these monosomies in our cohort is explainable, since we predominantly assessed t-AML (17 out of 18 cases). Indeed, it is known that monosomy 5 or 7 is more often associated with a disease that is preceded by t-MDS[15,25].

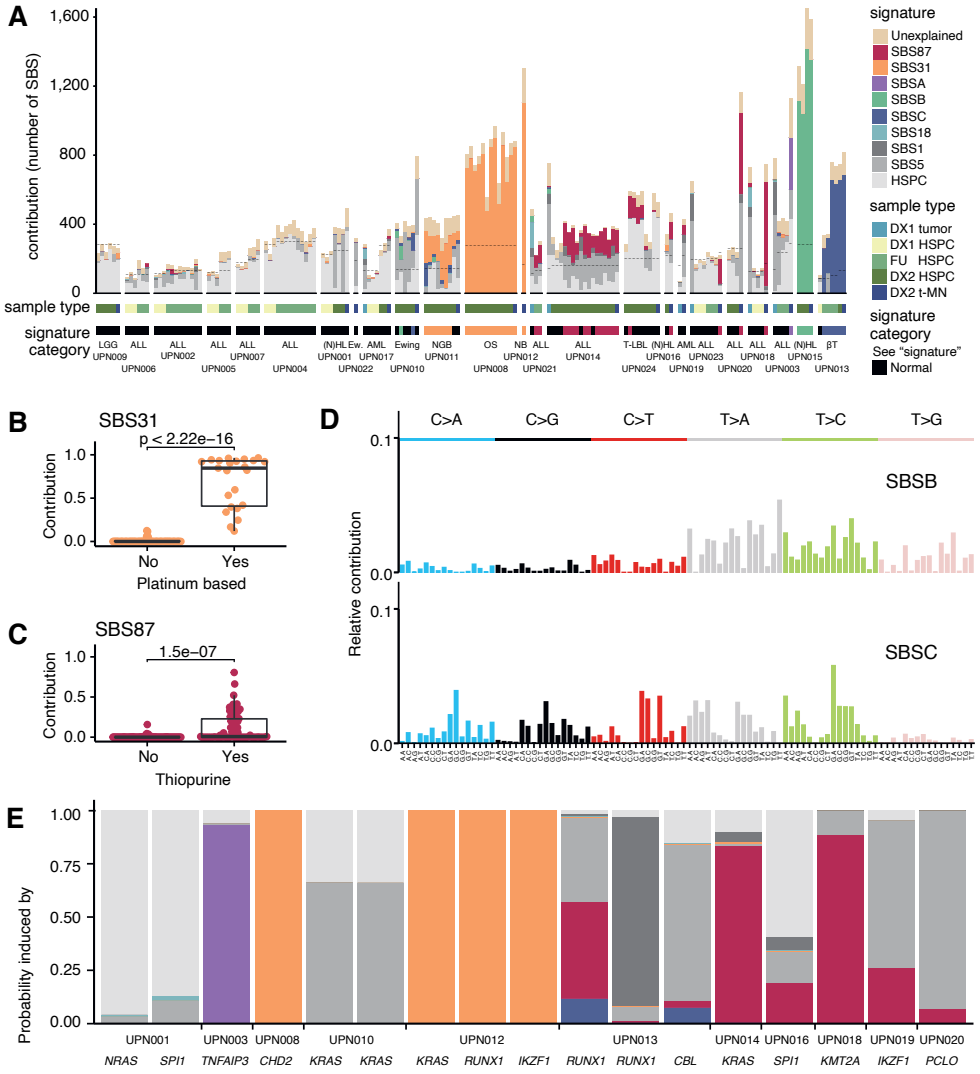


Figure 2. Thiopurines and platinum-based compounds have direct mutagenic effects in posttreatment HSPCs.

A) The contribution of each SBS signature to each sample, as obtained after refitting of signatures that were extracted by non-negative matrix factorization. The first row of bars below the plot indicates the timing of each sample. The second row indicates which treatment signature had more than 20% contribution in each sample, these HSPCs were termed t-HSPC. HSPCs without more 20% contribution of any of these signatures were termed n-HSPC. ALL, acute lymphoblastic leukemia; βT, betathalassemia; Ew., Ewing; LGG, low-grade glioma; NB, neuroblastoma; NGB, neuroganglioblastoma; (N)HL, (non-) Hodgkin lymphoma; OS, osteosarcoma; SBS, single base substitutions; T-LBL, T-cell lymphoblastic lymphoma. **B)** The contribution of SBS31 to samples treated or not treated by platinum-based drugs (two-sided Wilcox test). **C)** The contribution of SBS87 to samples treated or not treated by thiopurines (two-sided Wilcox test). **D)** The 96-trinucleotide single base substitution profiles of SBSB and SBSC. **E)** The probability that different driver mutations were caused by treatment-related or clock-like signatures.

Chemotherapy exposure increases the mutation burden in vivo

During healthy life, HSPCs accumulate mutations in a linear fashion with a rate of 14 to 15 single base substitutions and about one indel per year[18,19]. We compared the mutation burden of the pre- and posttreatment HSPCs to this baseline to correct for age-related mutation accumulation (**Fig. 1E,F**, **Fig. S2D**). The pretreatment HSPCs showed a mutation burden that was similar to this baseline. In contrast, the posttreatment HSPCs showed an increased number of single base substitutions and indels, corresponding to mutational ages 26 up to 94 years for single base substitutions [mean increase of 16; 95% confidence interval (CI), 13-19] and 90 years for indels [mean increase of 16; 95% CI, 12-19] (**Fig. 1E, F**). Furthermore, the t-MN blasts showed an increased mutation load compared with the baseline, but also compared to matched posttreatment HSPCs (**Fig. 1G**). This latter observation indicates that at the time of t-MN initiation, the leukemic cell of origin suffered more from mutagenesis than the average exposed HSPC. Nonetheless, similar to previous reports, the mutation load in t-AML blasts was only slightly higher than de novo pediatric AML and this difference was not significant in our analysis ($P = 0.053$, **Fig. 1H,I**)[13,16,19]. Together, our data suggest that for pediatric AML to arise, albeit related to treatment or naïve, a minimal mutation burden seems to be required, which is higher than the average number of somatic mutations observed within the healthy HSPC population.

Mutational signatures induced by chemotherapy exposure

To identify the processes that underlie the increased mutation burden in posttreatment HSPCs and t-MN blasts, we analyzed mutation spectra and signature contributions[27]. As expected, the mutation spectra of pretreatment HSPCs were similar to treatment-naïve cells and could be predominantly explained by the HSPC signature[18,20,28] and to a lesser extent Catalogue Of Somatic Mutations in Cancer (COSMIC) signatures SBS1 and SBS5[29] (**Fig. 2A**). Indeed, the contribution of these three clock-like signatures increases with age in healthy HSPCs[18,19]. In the posttreatment HSPCs and t-MNs, we additionally identified SBS31 and SBS87, which are caused by platinum-based drugs and thiopurines, respectively[7,30]. SBS31 and DBS5, also caused by platinum-based drugs[30,] were present in all cells of platinum-exposed patients (N=3; **Fig. 2A,B, Fig. S3A,B**). In t-MN patients that received thiopurine therapy (N=9), all but one t-MN genome harbored SBS87 mutations (**Fig. 2A,C**). In contrast to platinum-based exposure, only some posttreatment HSPCs displayed SBS87, indicating that thiopurine exposure is not always mutagenic to all cells. Furthermore, we identified three novel signatures that likely represent distinct mutational processes, as they could not be accurately decomposed by three or fewer existing signatures (**Fig. 2A,D, Fig. S3C-S3F**)[31]. Of these, SBSA was recently shown to be caused by the antiviral nucleoside analogue ganciclovir[32], and was present in the previously reported t-MN of patient UPN003 after exposure (**Fig. S4A,B**). The other two signatures (SBSB and SBSC) were observed in all posttreatment samples of single patients (UPN015 and UPN013, respectively). SBSB was associated with single thymidine deletions

at short T-repeats, while SBSC samples harbored large deletions at locations with microhomology (**Fig. S5A**). The T>N and C>T changes that contributed to SBSC displayed a wide sequence context preference for guanine at the -2 position of the mutated base (**Fig. S5B**). Although for patient UPN015 no treatment data was available, patient UPN013 received the alkylating drugs thiotepa and treosulfan as part of a conditioning treatment for hematopoietic stem cell transplantation. This patient was treated with multiple hematopoietic stem cell transplantations - which partly failed - to treat beta-thalassemia (**Table S1**). Posttreatment HSPCs of this patient displayed an increasing contribution of SBSC mutations after each consecutive round of transplantation (**Fig. 2A**). This suggests a causative role for the conditioning treatments in inducing SBSC mutations. Indeed, a pretreatment HSPC of this patient did not harbor SBSC mutations. We treated cord blood-derived HSPCs with thiotepa and treosulfan *in vitro*, after which we performed WGS on exposed cells, as described previously[33]. The resulting profiles showed similarities to SBSC but could not fully explain the signature (**Fig. S5C,D**).

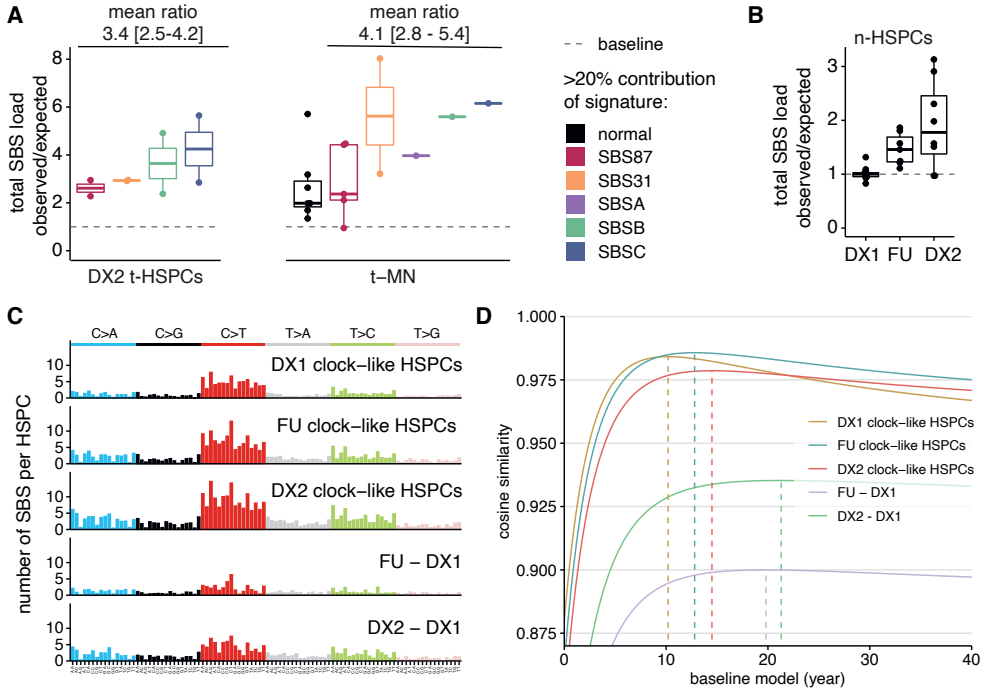


Figure 3. Chemotherapies induce a direct and indirect increase of mutations
A) The mutation load of single base substitutions (SBS) for DX2 HSPCs and t-MN grouped by signature category if 20% or more mutations can be explained by signature SBS87, SBS31, SBSA, SBSB or SBSC. **B)** The baseline-normalized mutation load of the DX1, FU and DX2 n-HSPCs. **C)** The average 96-trinucleotide SBS profile of n-HSPCs per timepoint. "FU – DX1" depicts the average profile of the FU n-HSPCs with the DX1 n-HSPCs subtracted. "DX2 – DX1" is similar but for DX2 with DX1 subtracted. **D)** the cosine similarity of each profile in C with a model of the baseline. Each dotted line indicates the age of the baseline at which the correlation of that profile is most similar to the model.

Although the mutation burden of most assessed t-MN cases was higher than the healthy baseline (**Fig. 1H**), only half of these cases displayed evidence of direct mutagenesis by chemotherapeutic compounds. To study the likelihood that leukemic driver mutations were directly caused by such compounds, we used a previously established method to calculate the probability that a certain mutation can be attributed to a signature[34,35]. This analysis showed that only seven of 17 identified driving single base substitutions had high probability (>70%) to be attributed to a treatment-related signature, whereas nine drivers were best explained by clock-like signatures (**Fig. 2E**). Therefore, in contrast to the identified fusion genes, which were likely the result of TOP2i, additional driver mutations in t-MN could only partly be explained by direct mutagenic consequences of chemotherapy treatment.

Direct and indirect induction of mutations after chemotherapy

Although we identified chemotherapy-associated signatures in a subset of posttreatment HSPCs and t-MN samples, most samples showed mutation spectra similar to pretreatment HSPCs, which could be fully explained by clock-like signatures (cosine similarity 0.97; **Fig. S6A**,**B**). Given this difference, we defined two categories: t-HSCPs, which have a spectrum that is >20% explained by contribution of a treatment-related signature, and n-HSPCs, which have a normal spectrum that is similar to healthy cells. Surprisingly, the mutation burden not only in t-HSPCs, but also in n-HSPCs was elevated compared with age-matched treatment-naïve HSPCs (**Fig. 3A,B**). The mutation load increase of n-HSPCs at FU and DX2 was 1.47-fold (95% CI, 1.25-1.69) and 1.91-fold (1.34, 2.47) compared with the baseline, respectively. Only the posttreatment HSPCs of patients UPN017 and UPN020 did not display elevated mutation burdens compared with the healthy baseline (**Fig. 2A**, **Fig. S6C**). The absence of an increased mutation burden in the HSPCs of these patients was not explained by a lack of exposure to chemotherapy, as UPN017 was treated for pAML and UPN020 for acute lymphoblastic leukemia (ALL). Indeed, the t-MN blasts of UPN020 harbored SBS87 mutations as well as an increased number of indels (**Fig. 2A**, **Fig. S3A**). Interestingly, both patients showed a longer latency time to t-MN development compared with the rest of the cohort. The posttreatment HSPCs of UPN017 were isolated with the longest latency after end of treatment (9.3 years vs 0.2-4.7 years in the rest of the cohort). The end of treatment for UPN020 was unknown, but the latency time between primary cancer and t-MN was 5.8 years (**Table S1**). These observations may suggest that years after treatment, HSPCs with a lower mutation load may preferentially contribute to hematopoiesis, similar to what has been reported in bronchial cells of ex-smokers[36]. Indeed, the risk for developing t-MN after treatment of a solid tumor peaks at 2 years after treatment and has been reported to return to a baseline population risk in 10-15 years[37].

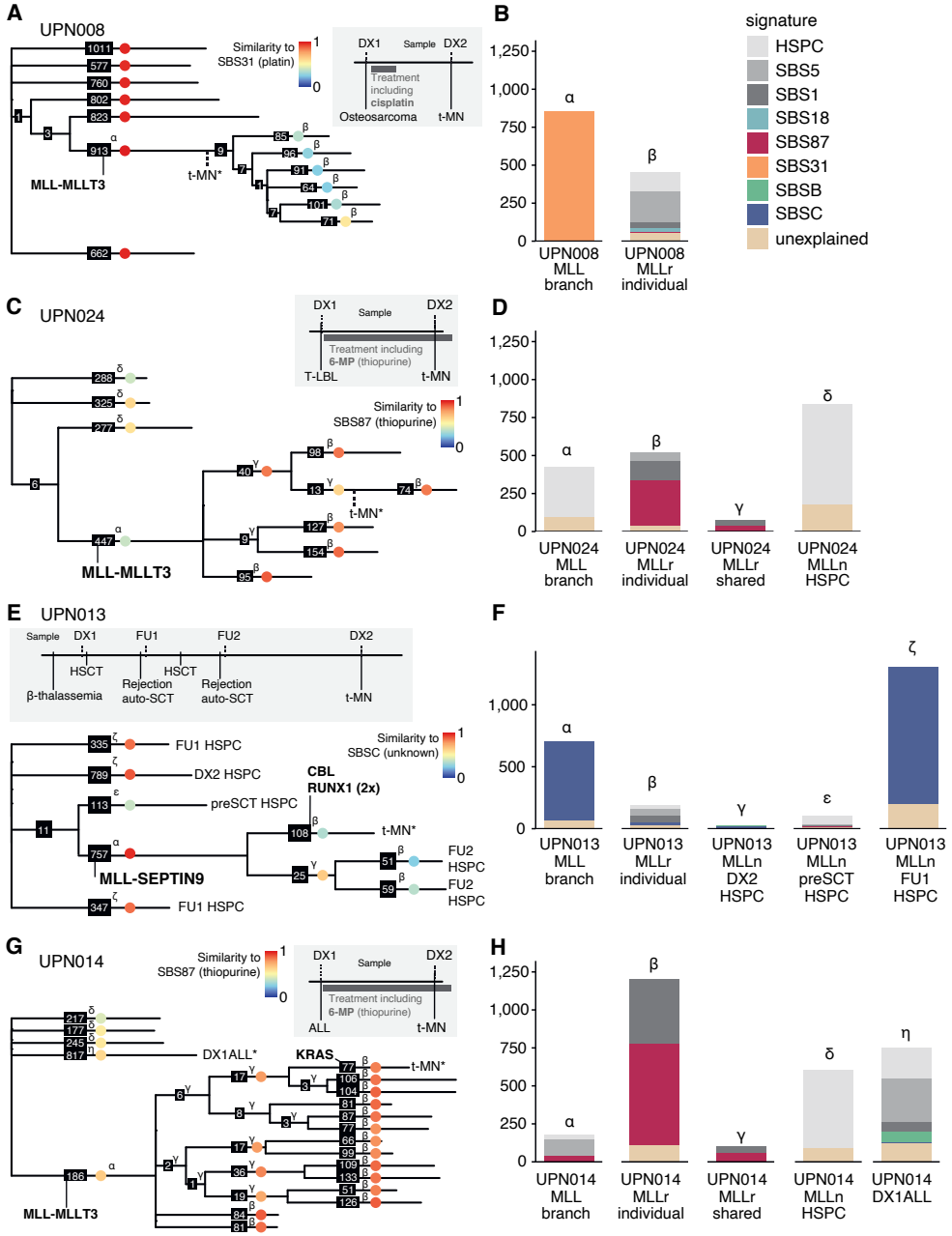


Figure 4. MLLr HSPCs in MLLr t-MN patients

A) Phylogenetic tree of DX2 HSPCs and bulk t-MN blasts of patient UPN008. Branches without a label represent DX2 HSPCs. Colored dots indicate the similarity of the 96-trinucleotide profile of each branch with more than 10 mutations with SBS31. The numbers indicate the number of single base substitutions (SBS) and indels in that branch. Sample marked with an asterisk is the only one that harbored blast markers. Top right, schematic overview of the disease, treatment, and sample collection time line for this patient. *The legend continues on the next page.*

In this patient, the t-MN blasts had no unique mutations. platin, cisplatin. **B)** The signature contribution of the mutations in the corresponding lineage trees on the left. **C)** Similar to A but for patient UPN024; similarity to SBS87 is depicted in the colored dots. Sample marked with an asterisk is the only one that harbored blast markers. **D)** Similar to B but for patient UPN024. **E)** Similar to A but for patient UPN013; similarity to SBSC is depicted in the colored dots. **F)** Similar to B but for UPN013. **G)** Similar to A but for patient UPN014 and similarity to SBS87 is depicted in the colored dots. Samples marked with an asterisk are the only one that harbored blast markers. All other samples were sorted on HSPC markers. **H),** Similar to B but for UPN014. α, the MLL rearrangement-containing branch; β, the aggregate of the unique mutations in the MLLr samples; γ, the shared mutations of the MLLr samples after the MLL-containing branch; δ, MLL-normal DX2 HSPC, ε, MLL-normal pre-SCT HSPCs; ζ, MLL-normal FU1 and DX2 HSPCs; η, the primary ALL.

As SBS1 and SBS5 are mainly active during fetal hematopoietic development and the HSPC signature postnatally, the mutation spectrum of healthy HSPCs changes during the first years of life[18,19,29,38,39]. To estimate if similar mutations accumulated during treatment as during aging, we determined the similarity between the mean SBS profile of the n-HSPCs to the mutation accumulation baseline in healthy HSPCs (**Fig. 3C,D**). Compared to the profile of DX1 HSPCs, the n-HSPC profile at time of FU1 and DX2 was more similar to the profile of older, healthy HSPCs. This observation was even more apparent for the profile of additional mutations in posttreatment HSPCs, showing that not only the mutation burden, but also the mutation spectra of these HSPCs are similar to those of healthy individuals of an older age. These analyses suggest that chemotherapy exposure can lead to an increased mutational age of normal HSPCs *in vivo*. Importantly, this indirect mutagenic effect of chemotherapy exposure may contribute to the accumulation of t-MN driver mutations (**Fig. 2E**).

Phylogenetic history of t-MN

To time the mutagenic effect of chemotherapy during t-MN development, we delineated the phylogenetic history using somatic mutations that were shared between cells of the same patient[18,20,40]. Although most posttreatment HSPCs did not harbor cancer driver mutations, we identified some HSPCs with structural rearrangements. Three HSPCs in two patients harbored genetic fusion genes that were not shared by the t-MN (**Fig. S6D**) and importantly, in four t-MN cases we identified phenotypically HSPC-like cells, which shared the MLL rearrangement with the t-MN blasts (**Fig. 4**). For patient UPN008, all 12 posttreatment HSPC clones and the t-MN blasts predominantly harbored SBS31 mutations (**Fig. 4A,B**). Indeed, this patient was initially treated for osteosarcoma, which included platinum-based drugs (**Table S1**). We identified six MLLr HSPCs that also shared all the clonal somatic mutations present in the t-MN and additionally harbored 71 to 101 unique mutations each (β branches), of which some were sub-clonally present in the bulk t-MN sample (**Fig. 4A**). These unique mutations were predominantly attributed to SBS5 and SBS1, with low similarity to SBS31 (**Fig. 4B**). This indicated that they were acquired after cisplatin exposure and thus that the t-MN expanded after cisplatin treatment. Based on the mutation data, these MLLr "HSPCs" were part of the leukemic blast population despite their HSPC-like phenotype and lack

of CD33 expression, which was the blast-defining marker that was used to sort the blast population. This phenomenon resembles a previous report of MLLr infant ALL, where a HSPC-like blast population was reported that lacked expression of the characteristic B-cell marker CD19[41]. We observed this phenomenon in patient UPN024, treated for T-cell lymphoblastic lymphoma (T-LBL), where one HSPC-like cell shared all mutations with the t-MN, thus genetically characterizing as a leukemic cell. Interestingly, similar to UPN008, we found an MLL rearrangement as sole genetic driver of the t-MN of UPN024. This rearrangement was shared by an additional four HSPC-like cells that did not harbor all other t-MN mutations in their genomes (**Fig. 4C,D**). This observation suggests that additional nongenetic hits are required for full malignant transformation.

In patients UPN013 and UPN014 (**Fig. 4E-H**), the t-MN had additional driver mutations that were not shared with the MLLr HSPCs (*RUNX1*/*CBL* and *KRAS*^{G12A}, respectively), indicating these were preleukemic cells that separated from the t-MN lineage before the leukemic cell of origin started expanding. In patient UPN013, treated for beta-thalassemia (see above), the mutations shared between MLLr HSPCs and t-MN blasts were all attributed to SBSC [**Fig. 4E** (α branch)]. In contrast, the clone-specific mutations (β branches) were mostly attributed to SBS1 and SBS5, indicating the MLLr HSPCs separated from the t-MN lineage at the end of mutagen exposure, similar to UPN008 (**Fig. 4F**). In patient UPN014, who developed t-MN after a first diagnosis of ALL, the 13 MLLr HSPCs shared 186 single base substitutions/indels with the t-MN [**Fig. 4G** (α branch)]. The MLLr branch had an estimated length of 8.0 years (Methods), while the patient was 7.1 years old at t-MN diagnosis. In this branch, 37 mutations could be attributed to SBS87 (**Fig. 4H**). These observations suggest that the first detectable division of the MLLr cell occurred during thiopurine-exposure. As the timing of MLL rearrangement within this branch cannot be further determined, it is unclear if this initial t-MN driver event in this patient was acquired before or after initiation of treatment. The unique mutations in the t-MN and MLLr HSPCs (β branches) were predominantly attributed to SBS87 and SBS1, whereas three posttreatment non-MLLr HSPCs only showed the HSPC signature (δ branches), similar to what was observed in patient UPN024 (**Fig. 4D,H**). The lack of SBS87 mutations in these latter cells is likely explained by the quiescent state of normal HSPCs[42] and the dependency of thiopurine-induced mutagenesis on replication[43]. In contrast, the t-MN and MLLr HSPCs of UPN014 and UPN024 did harbor SBS87 mutations (**Fig. 2A**; **Fig. S6E**), suggesting that their predecessors were replicating during thiopurine treatment. This idea is further supported by the presence of SBS1 mutations in these cells, which is a signature that has been associated with cell division[29] (**Fig. 4D,H**). The data together suggest that cell division, which may have been propagated by MLL rearrangement[44], during thiopurine therapy results in the accumulation of passenger and driver mutations (**Fig. 2E**).

Discussion

Previous studies have reported the mutational effects of chemotherapy exposure in cell culture systems, metastasized cancers, or colonic crypts[3,6,20]. Here, we report the first systematic analysis of chemotherapy-associated mutation accumulation in normal blood cells of pediatric cancer patients. Our t-MN patient cohort includes a large variety in clinical characteristics, such as age at first cancer diagnosis, type of first cancer and treatment regimen. Due to the relative rarity of the disease, and thus limited availability of samples in individual centers and even countries, international collaboration is essential to build larger cohorts and learning more about the specific mechanisms behind pediatric t-MN development. Despite these limitations, we could systematically confirm hypotheses that chemotherapy mutates normal HSPCs[5]. Furthermore, t-MN blasts showed an even higher mutation load compared with DX2 HSPCs, resulting in similar mutation numbers to pAML. Phylogenetic analyses allowed us to time t-MN, which elucidated different timing of t-MN expansion between treatments. Lastly, patients who had the longest latency to t-MN development showed a lower mutation load, which could mean that in those patients HSPCs with fewer mutations took over the blood system.

To elaborate, we found that posttreatment HSPCs of childhood cancer patients harbored a mutation burden comparable with HSPCs of adults. Collectively, these HSPCs showed a mean increase of 16 years of mutational age with excesses up to 80 years. In some patients, this increment in mutation load could be attributed to direct mutagenesis by thiopurines or platinum-based drugs, as reflected by mutational signatures (SBS87 and SBS31, respectively). Where in previous literature of pediatric t-MN both SBS31 and SBS35 have been linked to platinum therapy[13,16], we here mainly identified SBS31. Although the originally extracted non-negative matrix factorization (NMF) signatures did show characteristics of SBS35, subsequent refitting revealed a larger cosine similarity with SBS31 (0.98). In contrast to these clear therapy-related signatures in some cells, we show that in most HSPCs the increase in mutation burden upon chemotherapy exposure could not be explained by direct chemotherapeutic drug-induced mutagenesis. Their mutational profiles were more similar to those of older, healthy individuals, indicating that treatment predominantly causes indirect mutagenesis in exposed HSPCs. This indirect mutagenesis could be attributed to SBS5 and HSPC signatures, but not to SBS1, which is also observed in treatment naïve HSPCs during healthy aging[18,20] and in line with the predominant quiescent state of these cells after birth[42]. Moreover, hematopoietic stem cell transplantation in leukemia patients does not result in increased HSPC mutation loads[32], making bone marrow repopulation after therapy an unlikely cause for the observed increase in mutation load. The lack of direct chemotherapy-associated signatures in exposed HSPCs corresponds to recent data on environmental carcinogens in various mouse tumors[31], suggesting similar mechanisms may be active in other tissues. Thus, the origin for this indirect mutagenesis may be replicative stress[45] or stress-induced mutagenesis[46].

We found that direct mutagenesis by chemotherapeutic drugs may have varying dependencies. Whereas thiopurine-induced mutagenesis critically depended on cell division, platinum-based drugs were mutagenic to all assessed cells of exposed patients. The cisplatin-induced mutations in normal HSPCs support an earlier hypothesis that nonmalignant cells are first damaged by chemotherapy before developing into t-MN[5]. Our observations that most HSPCs do not harbor SBS87 mutations after thiopurine treatment are in line with previous literature reporting on the lack of 5-FU-related mutations in exposed t-MN cases, which was believed to be caused by quiescence of normal cells at time of treatment[5]. Therefore, our data suggest that for the mutagenic action of cisplatin, cell proliferation is not required. Indeed, cisplatin covalently binds to base residues in double-stranded DNA and was previously reported in WGS data to induce mutations in all exposed t-MNs[5,13]. In addition, our findings imply that MLLr cells associated with DNA cross-linking treatment can only divide after the end of exposure, whereas MLLr cells can start dividing during treatment with the thiopurine base analogues.

In four cases, we observed that HSPCs acquired an MLL fusion (either before or during treatment) and gave rise to a pool of (pre-)leukemic, HSPC-like cells that started dividing during or directly after chemotherapy exposure. The additional driver mutations in two t-MN genomes indicate that the leukemic cell of origin started expanding and became dominant after the additional hit, which, according to our data, might be non-genetic events. We also identified two cases in which MLLr HSPCs were genetically indistinguishable from the t-MN, similar to reports of earlier described leukemic stem cells[47,48]. Unfortunately, for our patients, we did not have multiple longitudinal samples available from the period between first diagnosis and t-MN to further assess the clonal dynamics preceding t-MN. Deep sequencing of such retrospective samples could, in the future, shed more light on the timing and evolution of t-MN development[16]. Interestingly, the shared mutations of all MLLr cells in UPN008 do not harbor a driver mutation and were completely explained by SBS31 (platinum induced). This finding is in line with a previous report on three pediatric neuroblastoma patients, in whom CH, mainly consisting of platinum-induced mutations and no drivers, preceded the development of t-MN that arose after the acquisition of drivers[16]. This is in stark contrast to CH in adult cancer patients, in whom no platinum-induced mutations were found after treatment[5,49]. In conclusion, we showed that chemotherapy can be mutagenic in at least three ways: directly to all exposed cells by DNA cross-linking, directly to dividing cells by base analogue incorporation and indirectly by mimicking clock-like processes. All these mechanisms ultimately result in increased mutagenesis, which can contribute to t-MN development through induction of cancer driver mutations.

Methods

Patient samples

All bone marrow and peripheral blood samples were obtained via the biobank of the Princess Máxima Center for Pediatric Oncology with ethical approval under proposals OC2018-07, PMCLAB2018.026 and PMCLAB2020.151 in accordance with the Declaration of Helsinki. The mutational spectra from UPN003 were previously reported[32]. Patients' written informed consents were obtained by the University Medical Center Utrecht and the Princess Máxima Center. This study was approved by the Biobank Research Ethics committee of the University Medical Center Utrecht and the Biobank & Data Access Committee of the Princess Máxima Center. Five patients were first diagnosed with primary ALL and were in remission at the time that the follow-up (FU) sample was taken. These were UPN002, UPN004, UPN005, UPN006, UPN007. The other patients had diverse primary diagnoses and developed a t-MN (t-AML, N = 18, t-MDS, N = 1) later in life. One t-MN blasts sample (UPN021 DX2AML) was excluded as no clonal mutations were found in this sample, indicating that the sorted populations were not purely blasts. UPN009 received radiotherapy, but not chemotherapy as a treatment for the primary cancer and the DX2 samples of this patient were therefore excluded from mutation load analyses and posttreatment signature analyses.

FACS and HSPC culture

Bone marrow mononuclear cells were stained for fluorescence-activated cell sorting (FACS) after thawing. HSPCs were identified using the following surface markers: $Lin^-CD11c^-CD16^-CD34^+$, $CD38^-/CD45RA^-$ (**Fig. S1**). We defined (t-)MN blasts from both first and second diagnosis based on diagnostic immunophenotyping data if available. In most cases these blasts were CD33, CD38 and/or CD34 positive. ALL blasts from first diagnosis were defined based on diagnostic immunophenotyping data if available (mostly these were CD10, CD19 or CD7 positive).

Blasts and HSPCs were purified on a SH800S Cell Sorter (Sony, RRID:SCR_018066). First blasts were sorted in bulk for DNA isolation after which HSPCs were index sorted in a flat-bottom 384-well plate prepared with 75 uL HSPC culture medium per well. HSPC culture medium consisted of StemSpan SFEM medium (Stemcell technologies) supplemented with SCF (100 ng/mL), Flt3-ligand (100 ng/mL), IL-6 (20 ng/mL), IL-3 (10 ng/mL), TPO (50 ng/mL), UM729 (500 nM) and Stemregenin (750 nM).

For five samples (UPN001DX2 and UPN023DX1, UPN002DX1, UPN005DX1, UPN004DX1), the obtained sample was depleted for monocytic, pro T-cell or pro B-cell blasts (marked by anti-CD14, CD7 and CD10 respectively) using the EasySep anti-APC kit, following manufacturer's instructions. After blast deletion, we plated MSCs and sorted HSPCs following the same procedure as with all other samples.
HSPCs were cultured for 4 to 7 weeks at 37°C, 5% CO2 before collection.

Mesenchymal stromal cells (MSCs) were cultured from a fraction of bone marrow cells by plating bulk cells in 12-well culture dishes with DMEM-F12 medium (GIBCO), supplemented with 10% FBS. Medium was refreshed every other day to remove non-adherent cells and MSCs could be harvested when confluent, after approximately 2 to 3 weeks.

FACS antibodies

All antibodies were obtained from Biolegend, except for CD13 (Biosciences). Antibodies used for (t-)MN blast and HSPC populations: CD34-BV421 (clone 561, 1:20, RRID:AB_11147951), lineage (CD3/CD14/CD19/CD20/CD56)-FITC (clones UCHT1, HCD14, HIB19, 2H7, HCD56, 1:20, RRID:AB_10644012), 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), CD33-PE/Cy7 (clone WM53, 1:100, RRID:AB_2734264), CD49f-PE/Cy7 (clone GoH3, 1:100, RRID:AB_2561704), CD16-FITC (clone 3G8, 1:100, RRID:AB_314205), CD11c-FITC (clone 3.9, 1:20, RRID:AB_314173), CD123-Pe/Cy7 (clone 6H6, 1:100, RRID:AB_493577), CD13-PerCP/Cy5.5 (Biosciences, clone WM15, 1:20, RRID:AB_10645787), CD14-APC (HCD14, RRID:AB_830680). Additional antibodies used for depleting ALL blast populations: CD10-APC (clone HI10a, 1:100, RRID:AB_314920), CD7-APC (clone CD7-6B7, 1:100, RRID:AB_1877156).

Cord blood chemotherapy exposure

We used a previously established protocol[33] to treat cord blood-derived HSPCs with approximately IC50 concentrations of treosulfan and thiotepa (4uM and 12,5uM, respectively). Thiotepa treatment was combined with liver enzymes to support conversion to the active metabolites[6]. Used concentrations for these additional compounds were: 0,25% S9 fraction (Aroclor-1254-induced male Sprague Dawley rat liver), 3mM NADP (Sigma) and 15mM DL-isocitric acid trisodium salt hydrate (Sigma).

DNA isolation and WGS

DNA was isolated from cell pellets of blasts, MSCs and clonally expanded HSPCs using the DNeasy DNA Micro Kit (Qiagen), according to the instructions provided by the manufacturer. We modified this protocol slightly by adding 2µL RNase A (Qiagen) during the lysis step and eluting DNA in 50µL low EDTA TE buffer (10mM Tris, 0.1mM EDTA, G Biosciences).

For each sample, DNA libraries for Illumina sequencing were generated from at least 35 ng genomic DNA using standard protocols. The libraries were sequenced on Novaseq 6000 sequencers (RRID:SCR_016387; 2x150bp) at a depth of 15-30x. Two t-MN blast samples (UPN018 and UPN023) were sequenced to 90x coverage, as only a DNA pellet was available, and blast purity was 15% and 22% respectively. Reads were mapped to the human reference genome GRCh38 using the Burrows-Wheeler

Aligner v0.7.17 mapping tool with settings 'bwa mem –M –c100' [50]. Sambamba v0.6.8[51] was used to mark duplicate sequencing reads and GATK v.4.1.3.0[52] was used to perform base recalibration. See https://github.com/UMCUGenetics/NF-IAP for a full description and all code of the pipeline.

Mutation calling and filtering

Mutation calling and filtering was performed on multi-sample VCF files generated using HaplotypeCaller from GATK v.4.1.3.0. GATK's VariantFiltration was used for variant quality evaluation 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".

For two impure t-MN blast samples that were sequenced to 90x, the 'SNP_LowQualityDepth' filter was lowered to "QD<1".

Subsequently, SNPEffFilter[53], SNPSiftDbnsfp (database dbNSFP3.2a[54], GATK VariantAnnotator (database COSMIC v.89), and SNPSiftAnnotate (database GoNL release 5) were used for variant annotation.

Finally, to obtain catalogues of high-quality somatic mutation calls, we applied post-processing filtering steps, per patient, as described below (all scripts are available at: https://github.com/ToolsVanBox/SMuRF).

Briefly, only variants were considered that (i) were present on autosomal chromosomes; (ii) passed VariantFiltration with a GATK phred-scaled quality score ≥ 100; (iii) had a base coverage of at least 10X (30X samples) or 5X (15X samples) in the clonal and paired control sample; (iv) had a mapping quality (MQ) score of 60; (v) did no overlap with single nucleotide polymorphisms (SNPs) in the Single Nucleotide Polymorphism Database v146 and a panel of unmatched normal human MSC and fetal genomes (BED-file available upon request); (vi) had a GATK genotype score (GQ) of 99 (indel/sbs in clonal sample, or indel in paired control) or higher than 10 (sbs in paired control); (vii) had a variant allele frequency of >=0.3 (sbs/indel in 30x coverage sample, or indel in 15x sample) or >=0.15 (sbs in 15x coverage sample) or 0.07 (sbs/indels in the two 90x t-MN samples, see above) to exclude *in vitro* accumulated mutations; and (viii) did not have any evidence from a paired control sample (MSCs isolated from the same bone marrow) if available. For patients for which no matched MSC control was available, or when the control was contaminated with blast cells (UPN008/UPN014), instead of step (viii) a mutation was filtered out when it (a) was clonally present in all samples that passed QC for

that mutation, (b) was subclonally present in any sample or (c) was not confidently absent in at least one sample.
One HSPC clone (UPN013DX2 HSPC 1B23) was excluded as it did not match the fingerprint of the other samples of UPN013 and was likely a surviving donor cell from one of the two unsuccessful stem cell transplantations that were administered prior to sample collection. For UPN008 and UPN014, bulk MSC samples were excluded as they showed clear contamination with t-MN blasts as evidenced by reads supporting the t-MN driving fusion and subclonal presence of t-MN single-nucleotide variant/indel mutations.

Driver events

Single base substitutions or indels were considered driver events when they (i) had a MQ of 60, and a GQ of 10 or higher in both the (t-)MN and the paired control sample (if available) and minimal base coverage of 10x in both the (t-)MN and the paired control sample (if available); (ii) had a variant allele frequency higher than 0.3; (iii) were present in driver genes (either COSMIC Cancer Gene Consensus (version of 9/5/2019) or one of the frequently mutated genes in primary pediatric (t-)MN[23]; were (iv) a missense, frameshift, stop-gain, insertion or deletion; (v) had either a high or moderate expected effect as annotated by SnpEff; (vi) were not present in the Single Nucleotide Polymorphism Database v146 and a panel of unmatched normal human MSC and fetal genomes (BED-file available upon request); and (vii) had no evidence in the paired control samples (if available)[55].
Structural variant and chromosomal copy-number alteration calling was performed using the GRIDSS-purple-linx pipeline developed at the Hartwig Medical Foundation[56]. All structural variants were validated by hand using IGV[57] and false positive results were excluded. All whole chromosome duplications and deletions were considered driver events, as well as partial chromosomal (arm) gains and losses (reported as one category). Finally, translocation events resulting in fusion genes that involved at least one known AML driver gene (in this dataset *KMT2A* (MLL), *RUNX1*, *MECOM* or *MLLT10* were considered drivers.

Comparison with the baseline

When comparing mutation load, single base substitution and indel counts were normalized to GATK CallableLoci's CALLABLE length. The baseline data from previous publications was used[18,19]. As described before, a linear mixed-effects model was used to calculate the slope and intercept of the baseline while taking donor dependency into account using lme4 package in R[58]. Mutational ages were calculated based on the expected rate of mutation accumulation over time in HSPCs of healthy individuals, previously defined as baseline[18], as mutational_age = (number_of_mutations – baseline_intercept)/baseline_slope.

Mutational signature extraction and refitting

For the analysis of mutational patterns and signatures, the in-house developed R package MutationalPatterns v3.0.1[59] was used. For single base substitution analysis, first the 96-trinucleotide profiles per sample were extracted. Then, NMF was applied on this data combined with previously published mutational patterns of healthy tissues 18 to extract nine signatures ("extract_signatures" with options "rank = 9, nrun = 100"). These signatures were compared with the COSMIC mutational signature database v3.1[55] and a previously established clock-like signature in HSPCs (HSPC signature)[60]. Signatures with a cosine similarity of > 0.8 to one of the known signatures were replaced by that signature (1, 5, 18, 31, 87, HSPC). Of note, the signature replaced by SBS31 was very similar to this signature (cosine similarity of 0.98), but also had some characteristics of SBS35 (cosine similarity of 0.73). The other signatures were named SBSB and SBSC.

Only from one sample with more than 100 mutations, the cosine similarity between the reconstructed profile derived from these signatures and the original profile had a cosine similarity below 0.8 (UPN003 t-MN, 1,078 mutations, **Fig. S4A**). We have previously reported this sample, and showed that it had contribution of SBSA, a signature caused by ganciclovir. Therefore, we have added SBSA to the mutational signature repertoire, which resulted in a cosine similarity of UPN003 t-MN to 0.98 (**Fig. S4B**).

Fitting the signatures to the mutational profiles of samples

The resulting set of signatures was used to perform bootstrapped fitting using "fit_to_signatures_bootstrapped" with options "n_boots = 100, max_delta = 0.05". The bootstrap results were averaged per sample to get the contribution of each signature to the profile of each sample. Finally, per sample the number of sbs in the reconstructed 96-trinucleotide profile were subtracted from the original number of sbs and added as "unexplained" mutations.

The same steps were taken when refitting was performed when refitting on the per-branch profiles of phylogenetic trees (**Fig. 4**). The aggregate profiles (**Fig. 3C**) were acquired by first making an average profile of HSPCs per timepoint and then taking the mean from the resulting profiles per time point.

Determining signature categories

Cells with a contribution of more than 20% from signature SBSA, SBSB, SBSC, SBS31 or SBS87 were assigned as t-HSPC, and grouped to the corresponding signature category. Cells with more than 20% contribution of more than one of these signatures were grouped to the signature with the highest contribution. Finally, all remaining cells were assigned to the category n-HSPC, as most of their mutations could be attributed to SBS1, SBS5 and the HSPC signature. SBS18 was not a category, as in none of the HSPCs or t-MN samples did SBS18 have a contribution of 20% or more.

Genomic age estimation

The healthy 96-trinucleotide mutation data was obtained from previously sequenced HSPC clonal cultures[18,32]. For each trinucleotide category, a linear mixed effects model was applied to determine the age-related increase. Predictions for the 96-trinucleotide profiles of healthy aging were made for each timepoint at a resolution of 0.1 year. For each timepoint of our data set (DX1, FU, and DX2) the n-HSPC profiles were merged. Each resulting profile was compared to all baseline profiles using cosine similarity. The mutational age of each of the three n-HSPC profiles was set to age of the baseline profile with the highest cosine similarity.

Constructing phylogenetic trees

To construct phylogenetic trees, all samples from one patient were compared among one another. To obtain only high-confident mutations and to include mutations that arose during early development, filtering was slightly adjusted compared with previous analyses. If a control MSC sample was available, mutations that were subclonally (VAF<0.3) present in the control were considered. Mutations that were sub-clonally present in any other sample were filtered out. To still account for germline mutations, mutations that were clonally present in all samples were filtered out. In addition, all samples needed to have passed QC filters as described above (among others, sufficient coverage and mapping quality), not only the sample in which the mutation was found. Finally, for patient UPN013, 35 mutations were removed that were detected in all samples but the primary ALL, and that were present in locations with loss-of-heterozygosity in the ALL. All shared mutations were manually inspected in IGV, and false positive results were filtered out. A binary mutation table was constructed from the mutations that passed these criteria, and a tree was constructed using the ape v5.5 R package[61].
As filtering to obtain the tree is very strict, mutations that were filtered out due to failed QC in one or more samples were reconsidered. These mutations were added to a branch only if the VAF of that mutation was 0.15 for all samples in that branch and if the VAF was 0 for all samples not in the branch. In addition, mutations only found in 1 sample were only considered if that sample passed QC.

The mutations per branch were extracted using the binary mutation table and a cosine similarity to one of the NMF-extracted or COSMIC signatures was calculated. Then, the per-branch mutations were merged into categories, and refitting was performed on the resulting mutation catalogues as described above.

Potential impact of mutational signatures

Calculating the probability of a mutation being caused by the signatures that contributed to that sample was done similarly to that done by Morganella and colleagues[35]. 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.

Extended context

To determine the extended context of the mutations of post-treatment samples from patient UPN013, we extracted the -4/+4 context of each unique mutation in all samples. Next, we pulled all mutations for each of the six mutation types and plotted the sequence logos with the R package ggseqlogo v.0.1 [62].

Statistical analysis

Due to limited primary material availability, no sample-size calculation was performed. No randomization was performed. Regarding the included t-MN patients, all patients of which material at time of t-MN was available in the biobank were included in our study. For most samples at least three (with up to 16) HSPCs were sequenced. As specifically the t-MN material is scarce, we collected and processed all available samples to obtain this unique dataset. For the comparisons of mutation burden and signature contribution between groups, a two-sided Wilcox test was used.

Data and code availability

The datasets generated during this study are available at EGA (https://www.ebi.ac.uk/ega/), accession number EGA:EGAS00001005141. Most of the scripts used during this study are available at https://github.com/ToolsVanBox/ and in the MutationalPatterns R package (https://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html). Other scripts are available upon request.

Authors' disclosures

A.K.M. Rosendahl Huber, A.J.C.N. van Leeuwen, and R. van Boxtel report a patent for means and methods for assessing genotoxicity pending. R. van Boxtel reports grants from the European Research Council (ERC) and the Dutch Research Council (NOW) during the conduct of the study. No disclosures were reported by the other authors.

Author contributions

E.J.M. Bertrums: Data curation, formal analysis, validation, investigation, methodology, writing–original draft, project administration, writing–review and editing. A.K.M. Rosendahl Huber: Conceptualization, data curation, software, formal analysis, validation, investigation, methodology, writing–original draft, project administration, writing–review and editing. J.K. de Kanter: Data curation, software, formal analysis, investigation, visualization,
writing–original draft, writing–review and editing. A.M. Brandsma: Investigation, methodology. A.J.C.N. van Leeuwen: Investigation. M. Verheul: Investigation. M.M. van den Heuvel-Eibrink: Data curation, supervision, project administration. R. Oka: Software. M.J. van Roosmalen: Software. H.A. de Groot-Kruseman: Data curation.

C.M. Zwaan: Data curation, supervision, project administration. B.F. Goemans: Data curation, supervision, project administration. R. van Boxtel: Conceptualization, supervision, funding acquisition, writing–original draft, project administration, writing–review and editing.

Acknowledgements

This work was funded by an ERC consolidator grant from the European Research Council (ERC; no.864499) to R. van Boxtel. Additionally, this work was supported by the Oncode Institute, funding E.J.M. Bertrums., A.K.M. Rosendahl Huber, J.K. de Kanter, A.M. Brandsma, A.J.C.N. van Leeuwen, M. Verheul, R. Oka, M.J. van Roosmalen. and R. van Boxtel, and a VIDI grant from the Dutch Research Council (NOW; no. 016.Vidi.171.023) to R. van Boxtel that supports A.K.M. Rosendahl Huber. The authors thank the Hartwig Medical Foundation (Amsterdam, the Netherlands) for facilitating low-input WGS.

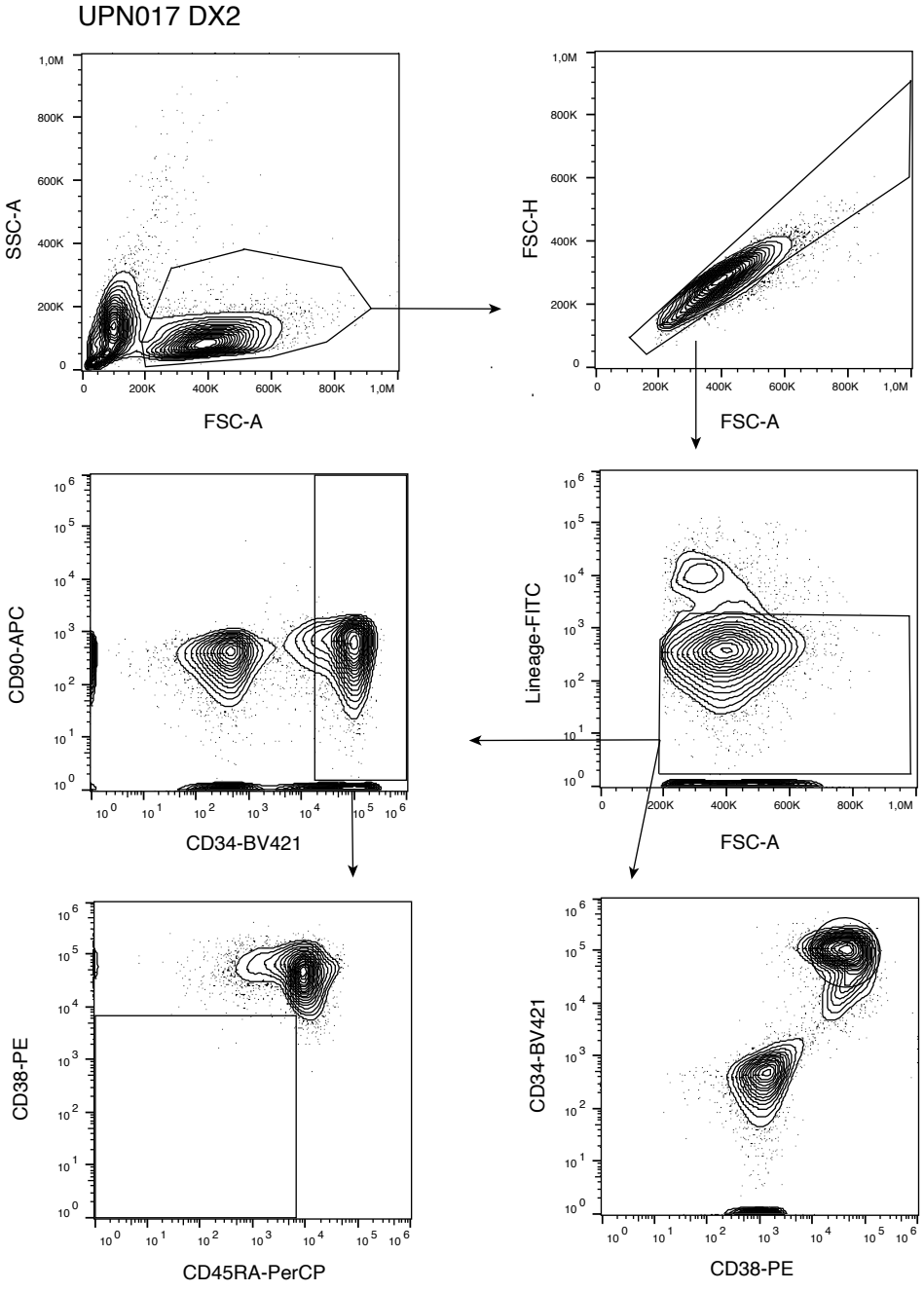
References

1. Chabner, B. A. & Roberts, T. G., Jr. Timeline: Chemotherapy and the war on cancer. Nat Rev Cancer 5, 65-72, doi:10.1038/nrc1529 (2005).
2. Hurley, L. H. DNA and its associated processes as targets for cancer therapy. Nat Rev Cancer 2, 188-200, doi:10.1038/nrc749 (2002).
3. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature 575, 210-216, doi:10.1038/s41586-019-1689-y (2019).
4. Pich, O. et al. The mutational footprints of cancer therapies. Nature genetics 51, 1732-1740, doi:10.1038/s41588-019-0525-5 (2019).
5. Pich, O. et al. The evolution of hematopoietic cells under cancer therapy. Nat Commun 12, 4803, doi:10.1038/s41467-021-24858-3 (2021).
6. Kucab, J. E. et al. A Compendium of Mutational Signatures of Environmental Agents. Cell 177, 821-836.e816, doi:10.1016/j.cell.2019.03.001 (2019).
7. Li, B. et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. Blood 135, 41-55, doi:10.1182/blood.2019002220 (2020).
8. Christensen, S. et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun 10, 4571, doi:10.1038/s41467-019-12594-8 (2019).
9. Crawford, J., Dale, D. C. & Lyman, G. H. Chemotherapy-induced neutropenia: risks, consequences, and new directions for its management. Cancer 100, 228-237, doi:10.1002/cncr.11882 (2004).
10. Robison, L. L. & Hudson, M. M. in Nat Rev Cancer Vol. 14 61-70 (2014).
11. Choi, D. K., Helenowski, I. & Hijiya, N. Secondary malignancies in pediatric cancer survivors: perspectives and review of the literature. International journal of cancer 135, 1764-1773, doi:10.1002/ijc.28991 (2014).
12. Cupit-Link, M. C. et al. Biology of premature ageing in survivors of cancer. ESMO Open 2, e000250, doi:10.1136/esmoopen-2017-000250 (2017).
13. Schwartz, J. R. et al. The acquisition of molecular drivers in pediatric therapy-related myeloid neoplasms. Nat Commun 12, 985, doi:10.1038/s41467-021-21255-8 (2021).
14. Bolton, K. L. et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nature genetics 52, 1219-1226, doi:10.1038/s41588-020-00710-0 (2020).
15. McNerney, M. E., Godley, L. A. & Le Beau, M. M. Therapy-related myeloid neoplasms: when genetics and environment collide. Nat Rev Cancer 17, 513-527, doi:10.1038/nrc.2017.60 (2017).
16. Coorens, T. H. H. et al. Clonal hematopoiesis and therapy-related myeloid neoplasms following neuroblastoma treatment. Blood 137, 2992-2997, doi:10.1182/blood.2020010150 (2021).
17. Takahashi, K. et al. Preleukaemic clonal haemopoiesis and risk of therapy-related myeloid neoplasms: a case-control study. The Lancet. Oncology 18, 100-111, doi:10.1016/s1470-2045(16)30626-x (2017).
18. Osorio, F. G. et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep 25, 2308-2316.e2304, doi:10.1016/j.celrep.2018.11.014 (2018).
19. Brandsma, A. M. et al. Mutation Signatures of Pediatric Acute Myeloid Leukemia and Normal Blood Progenitors Associated with Differential Patient Outcomes. Blood Cancer Discovery 2, 484-499,

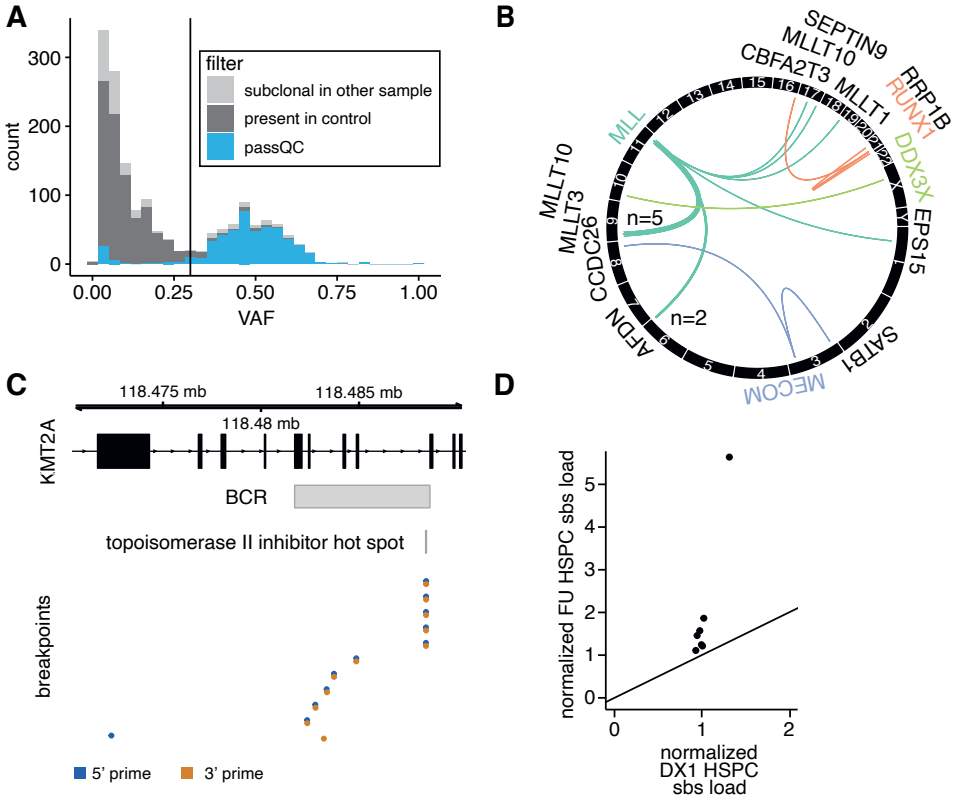
doi:10.1158/2643-3230.Bcd-21-0010 (2021).
20. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. Nature 561, 473-478, doi:10.1038/s41586-018-0497-0 (2018).
21. Le, H. et al. Rearrangements of the MLL gene are influenced by DNA secondary structure, potentially mediated by topoisomerase II binding. Genes, chromosomes & cancer 48, 806-815, doi:10.1002/gcc.20685 (2009).
22. Mirault, M. E., Boucher, P. & Tremblay, A. Nucleotide-resolution mapping of topoisomerase-mediated and apoptotic DNA strand scissions at or near an MLL translocation hotspot. American journal of human genetics 79, 779-791, doi:10.1086/507791 (2006).
23. Bolouri, H. et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nature medicine 24, 103-112, doi:10.1038/nm.4439 (2018).
24. Wong, T. N. et al. Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. Nature 518, 552-555, doi:10.1038/nature13968 (2015).
25. Pedersen-Bjergaard, J., Pedersen, M., Roulston, D. & Philip, P. Different genetic pathways in leukemogenesis for patients presenting with therapy-related myelodysplasia and therapy-related acute myeloid leukemia. Blood 86, 3542-3552 (1995).
26. Robinson, P. S. et al. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. Nature genetics 53, 1434-1442, doi:10.1038/s41588-021-00930-y (2021).
27. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. Cell Rep 3, 246-259, doi:10.1016/j.celrep.2012.12.008 (2013).
28. Maura, F. et al. A practical guide for mutational signature analysis in hematological malignancies. Nat Commun 10, 2969, doi:10.1038/s41467-019-11037-8 (2019).
29. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. Nature 578, 94-101, doi:10.1038/s41586-020-1943-3 (2020).
30. Boot, A. et al. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. Genome Res 28, 654-665, doi:10.1101/gr.230219.117 (2018).
31. Riva, L. et al. The mutational signature profile of known and suspected human carcinogens in mice. Nature genetics 52, 1189-1197, doi:10.1038/s41588-020-0692-4 (2020).
32. de Kanter, J. K. et al. Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. Cell stem cell 28, 1726-1739, doi:10.1016/j.stem.2021.07.012 (2021).
33. Rosendahl Huber, A. et al. Whole-genome sequencing and mutational analysis of human cord-blood derived stem and progenitor cells. STAR Protoc 3, 101361, doi:10.1016/j.xpro.2022.101361 (2022).
34. Brady, S. W. et al. The Clonal Evolution of Metastatic Osteosarcoma as Shaped by Cisplatin Treatment. Mol Cancer Res 17, 895-906, doi:10.1158/1541-7786.Mcr-18-0620 (2019).
35. Morganella, S. et al. The topography of mutational processes in breast cancer genomes. Nat Commun 7, 11383, doi:10.1038/ncomms11383 (2016).
36. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature 578, 266-272, doi:10.1038/s41586-020-1961-1 (2020).
37. Radivoyevitch, T. et al. Defining AML and MDS second cancer risk dynamics after diagnoses of first cancers treated or not with radiation. Leukemia 30, 285-294, doi:10.1038/leu.2015.258 (2016).
38. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. Nature genetics 47, 1402-1407, doi:10.1038/ng.3441 (2015).
39. Hasaart, K. A. L. et al. Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis. Scientific reports 10, 12991, doi:10.1038/s41598-020-69822-1 (2020).
40. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature 513, 422-425, doi:10.1038/nature13448 (2014).
41. Chen, C. et al. Single-cell multiomics reveals increased plasticity, resistant populations and stem-cell-like blasts in KMT2A-rearranged leukemia. Blood 139, 2198-2211, doi:10.1182/blood.2021013442 (2021).
42. Bowie, M. B. et al. Hematopoietic stem cells proliferate until after birth and show a reversible phase-specific engraftment defect. The Journal of clinical investigation 116, 2808-2816, doi:10.1172/jci28310 (2006).
43. Ling, Y. H., Nelson, J. A., Cheng, Y. C., Anderson, R. S. & Beattie, K. L. 2'-Deoxy-6-thioguanosine 5'-triphosphate as a substrate for purified human DNA polymerases and calf thymus terminal deoxynucleotidyltransferase in vitro. Mol Pharmacol 40, 508-514 (1991).
44. van der Linden, M. H. et al. MLL fusion-driven activation of CDK6 potentiates proliferation in MLL-rearranged infant ALL. Cell cycle (Georgetown, Tex.) 13, 834-844, doi:10.4161/cc.27757 (2014).
45. Flach, J. et al. Replication stress is a potent driver of functional decline in ageing haematopoietic stem cells. Nature 512, 198-202, doi:10.1038/nature13619 (2014).

46. Cipponi, A. et al. MTOR signaling orchestrates stress-induced mutagenesis, facilitating adaptive evolution in cancer. Science (New York, N.Y.) 368, 1127-1131, doi:10.1126/science.aau8768 (2020).
47. Chopra, M. & Bohlander, S. K. The cell of origin and the leukemia stem cell in acute myeloid leukemia. Genes, chromosomes & cancer 58, 850-858, doi:10.1002/gcc.22805 (2019).
48. Jordan, C. T. The leukemic stem cell. Best practice & research. Clinical haematology 20, 13-18, doi:10.1016/j.beha.2006.10.005 (2007).
49. Coombs, C. C. et al. Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. Cell stem cell 21, 374-382.e374, doi:10.1016/j.stem.2017.07.010 (2017).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
51. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2032-2034, doi:10.1093/bioinformatics/btv098 (2015).
52. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43, 491-498, doi:10.1038/ng.806 (2011).
53. Cingolani, P. et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front Genet 3, 35, doi:10.3389/fgene.2012.00035 (2012).
54. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80-92, doi:10.4161/fly.19695 (2012).
55. Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic acids research 47, D941-d947, doi:10.1093/nar/gky1015 (2019).
56. Cameron, D. L. et al. GRIDSS, PURPLE, LINX: unscrambling the tumor genome via integrated analysis of structural variation and copy number. bioRxiv, 781013 (2019).
57. Robinson, J. T. et al. Integrative genomics viewer. Nat Biotechnol 29, 24-26, doi:10.1038/nbt.1754 (2011).
58. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823 (2014).
59. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med 10, 33, doi:10.1186/s13073-018-0539-0 (2018).
60. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature 538, 260-264, doi:10.1038/nature19768 (2016).
61. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526-528, doi:10.1093/bioinformatics/bty633 (2019).
62. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics 33, 3645-3647, doi:10.1093/bioinformatics/btx469 (2017).

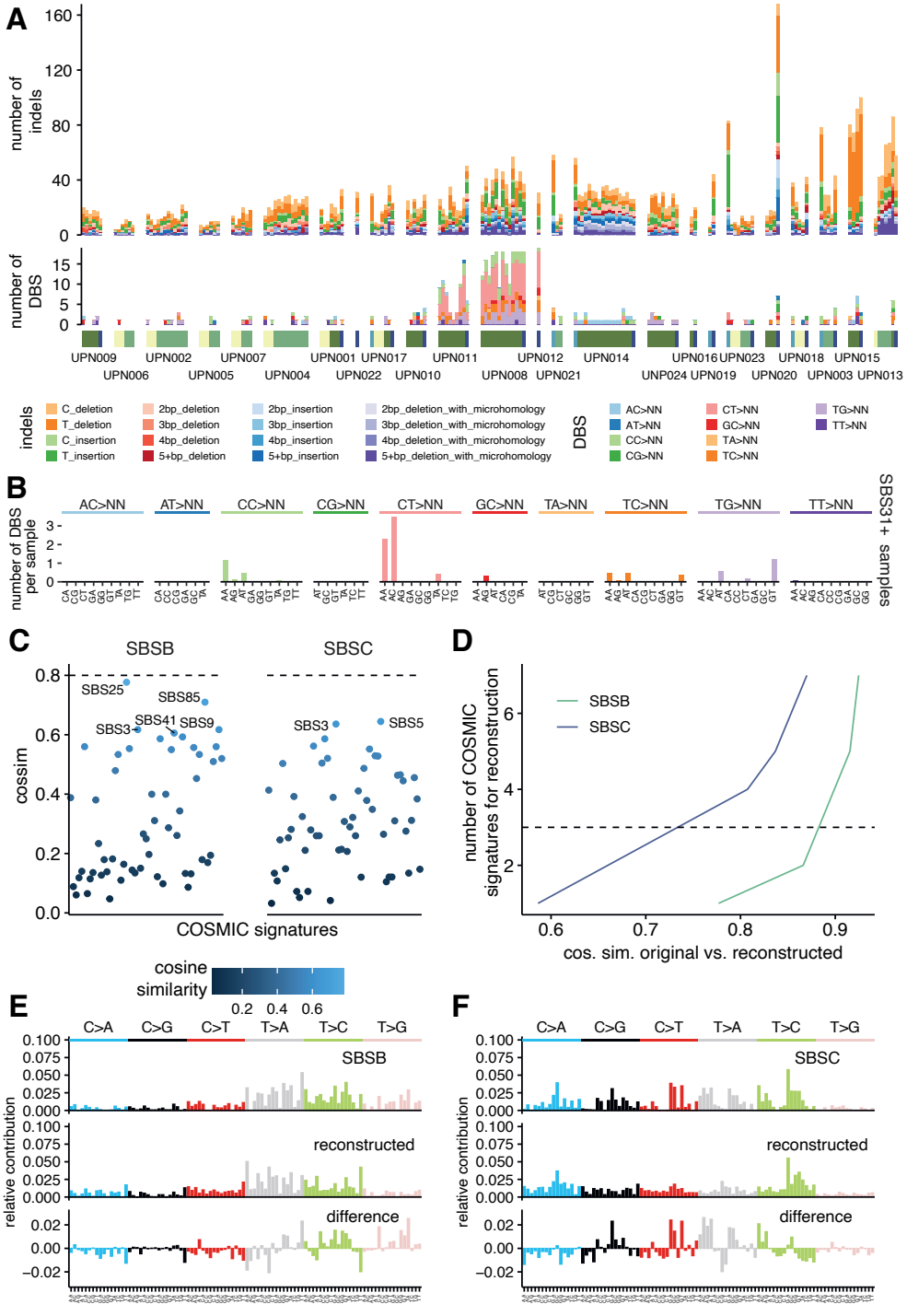
Supplementary Material



Supplementary Figure 1. Flow cytometry sorting strategy.
Representative fluorescence activated cell sorting (FACS) plot for purification of AML blasts (in UPN017 t-AML sample these are lin-CD34+CD38+) and single cell HSPCs (lin-CD34+CD38-CD45RA-).

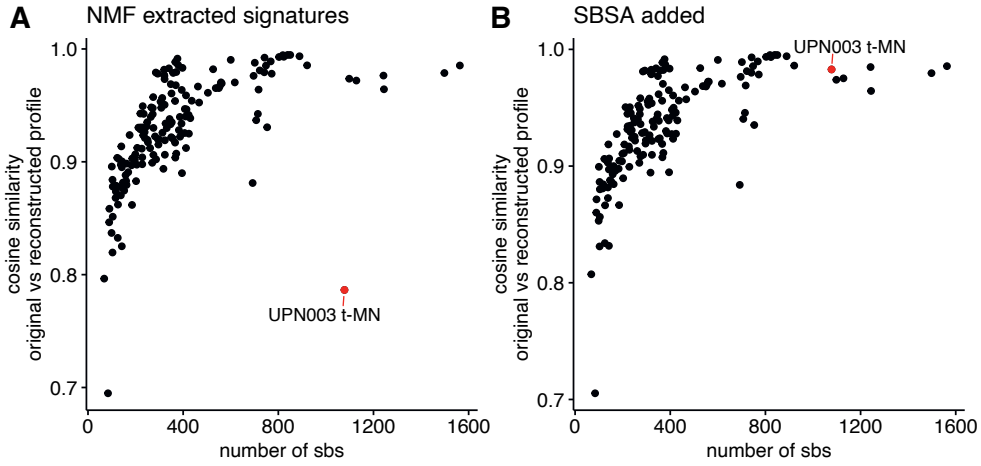


Supplementary Figure 2. Quality control (QC) and characteristics of the dataset.
A) Variant allele frequencies (VAF) of filtered and unfiltered mutations from a representative HSPC clone. Mutations that were subclonal in other samples, or present in the control sample were filtered out. Resulting "passQC" variants cluster around 0.5, and mutations with a VAF higher than 0.3 are selected as clonal. **B)** A circos plot of the structural variations of t-MN samples resulted in fusion driver genes. **C)** KMT2A (MLL) breakpoints from t-MN patients in our data set. Indicated above are the general KMT2A breakpoint cluster region (BCR) and the hot spot associated with topoisomerase II inhibitors (TOP2i). **D)** The single base substitution load in DX1 HSPCs compared to FU HSCPs of the same patient (both normalized to the baseline). Dots represent mean values of HSPCs per patient. The line has an intercept of 0 and a slope of 1, indicating the values at which DX1 and FU HSPCs would have the same, age-normalized mutation load.



Supplementary Figure 3. Indels, double base substitutions and extended context.
A) The number of small insertions and deletions (indels) and double base substitutions (DBS) of all samples in our data set. The sample type is depicted under the plot similarity to Figure 2A. **B)** The

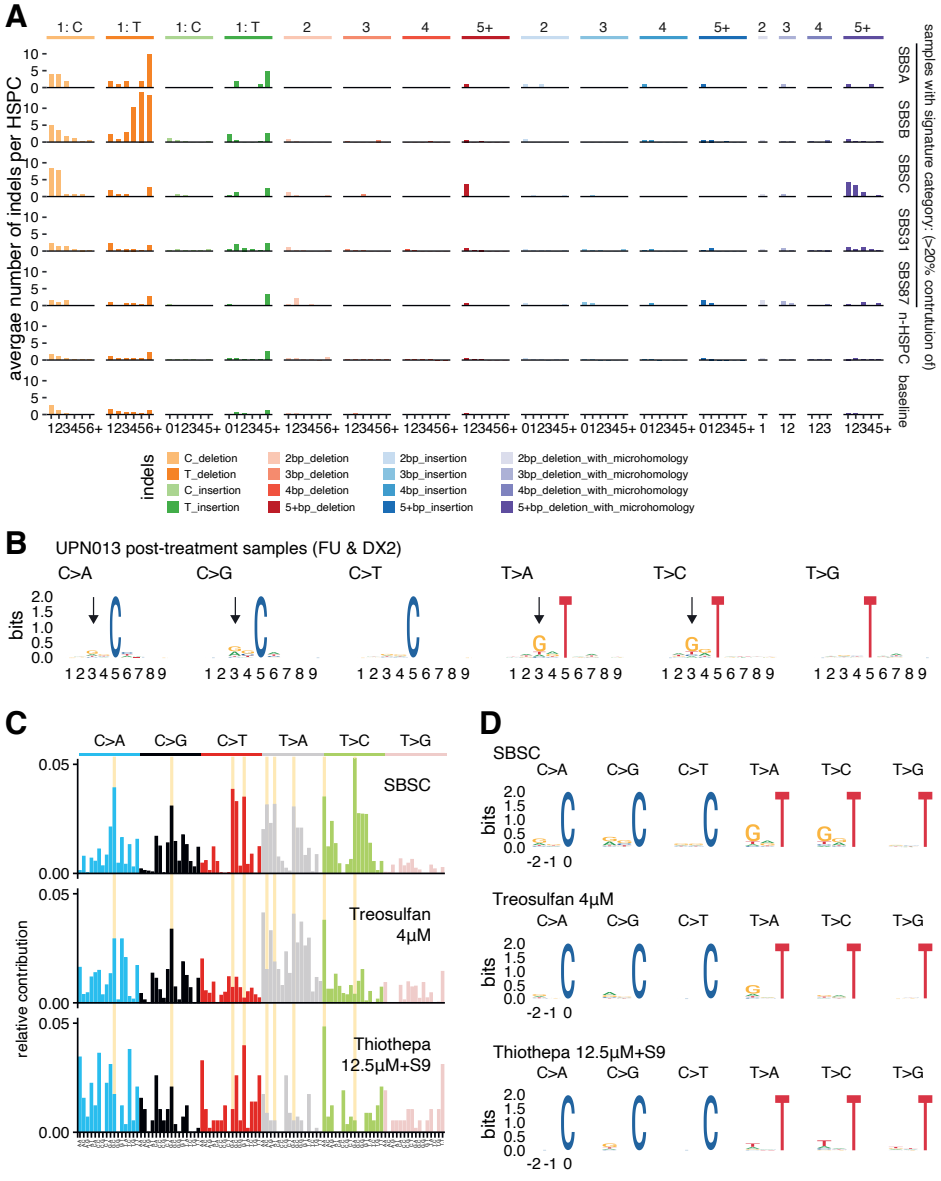
mean DBS profile of samples within the SBS31 signature category (>20% contribution of SBS31 to their 96-trinucleotide profile). This dbs profile is similar to COSMIC profile DBS5. **C)** Cosine similarity of SBSB and SBSC to COSMIC signatures. **D)** The cosine similarity of the reconstructed vs original profile of SBSB and SBSC when refitting a different number of the optimal COSMIC signatures to reconstruct the original profile. Both could not be reconstructed with three or less signatures to a cosine similarity higher than 0.9 **E)** The original 96-trinucleotide profile of SBSB, the reconstructed profile using the best max-delta cut-off (0.03, two signatures), and the difference of the two profiles. **F)** Similar to (E), but for SBSC with at max-delta=0.03 is reconstructed by five signatures.



Supplementary Figure 4. Cosine similarity of original versus reconstructed profiles before and after adding SBSA.

SBSB, SBSC and signatures with high similarity to SBS1, 5, 18, 31, 87 and the HSPC signature were extracted by NMF. Using these extracted signatures and their contribution to sample profiles, we reconstructed the sample profiles.

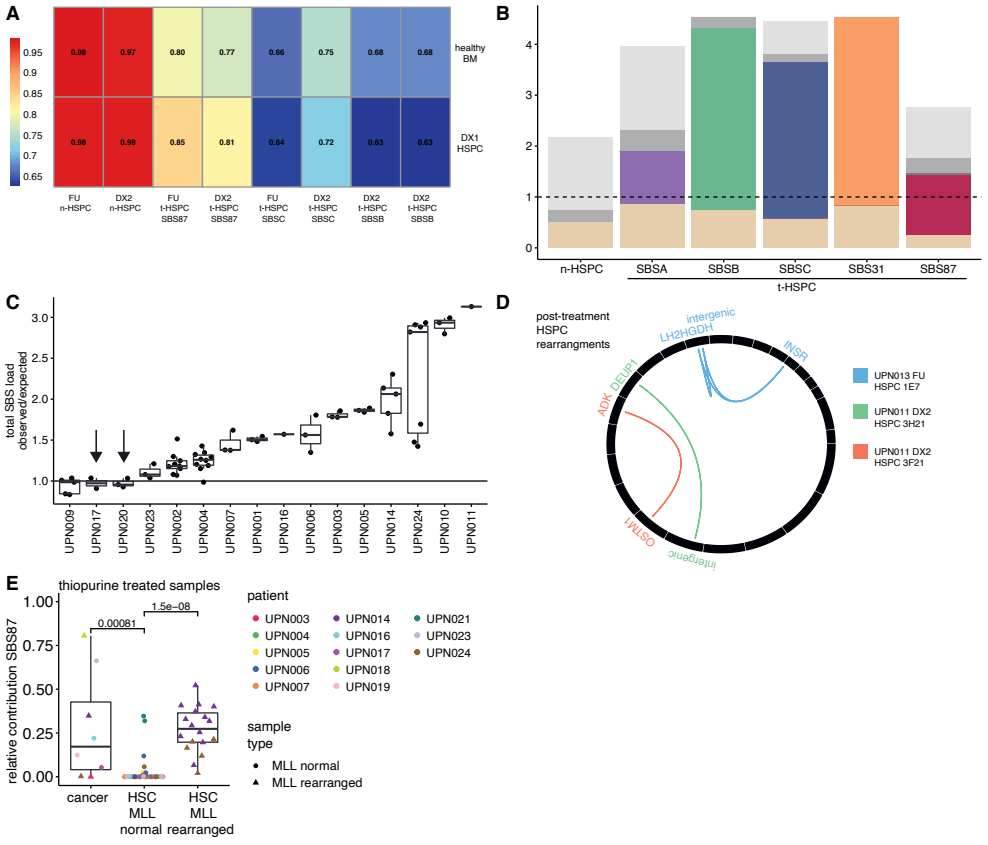
A) The cosine similarity of the reconstructed profiles and original profiles (y-axis) versus the number of mutations in each sample (x-axis). **B)** Similar to A, but the reconstructed profiles were created after adding SBSA to the signatures.



Supplementary Figure 5. SBSB and SBSC are distinct signatures.

A) The mean indel profile of the samples in each SBS signature category. **B)** The extended context of the six mutation types of post-treatment samples of patient UPN013. T>A, T>C, C>A and C>G mutations are enriched for guanines at position -2. SBSC and *in vitro* treatment of healthy cord blood cells with treosulfan or thiotepa. Cord blood cells from a healthy donor were treated *in vitro* with approximately IC50 concentrations for 72 hours, then were sorted as single cells and clonally expanded to harvest sufficient DNA for WGS. Thiotepa was converted into its active metabolite by adding S9 liver extract. **C)** From top to bottom, the 96-trinucleotide profiles of SBSC mutations, treosulfan mutations and thiotepa mutations. The *in vitro* profiles were corrected with mutations detected in untreated and S9-only treated cells respectively.

The -2, -1 and 0 (the mutated base) context preference of the 6 mutation types of the mutations in (C).



Supplementary Figure 6. HSPC profiles, mutation loads, structural variants and thiopurine mutations.

A) The cosine similarity of the posttreatment HSPCs of different signature categories (with >20% contribution of that signature) to the profile of the healthy baseline or the DX1 HSPCs.

B) The contributions of signatures to the mean baseline-corrected profile of posttreatment HSPCs of different signature categories.

C) The ratio of observed versus expected (based on the age of patient in combination with the baseline) mutations in post-treatment HSPCs of different patients. The arrows indicate patients that received chemotherapy but had no increased mutation load in their HSPCs. Patient UPN009 did not receive chemotherapy.

D) The structural variations found in HSPCs after treatment. None were found before treatment.

E) The relative contribution of SBS87, which is thiopurine-associated, to the mutation load of post-thiopurine-treatment t-AML samples, HSPCs without an MLL rearrangement and HSPCs with an MLL rearrangement. All samples without an MLL rearrangement are represented by a dot, all samples with an MLL rearrangement are represented by a triangle.

Supplementary Table 1. Patient and treatment information.

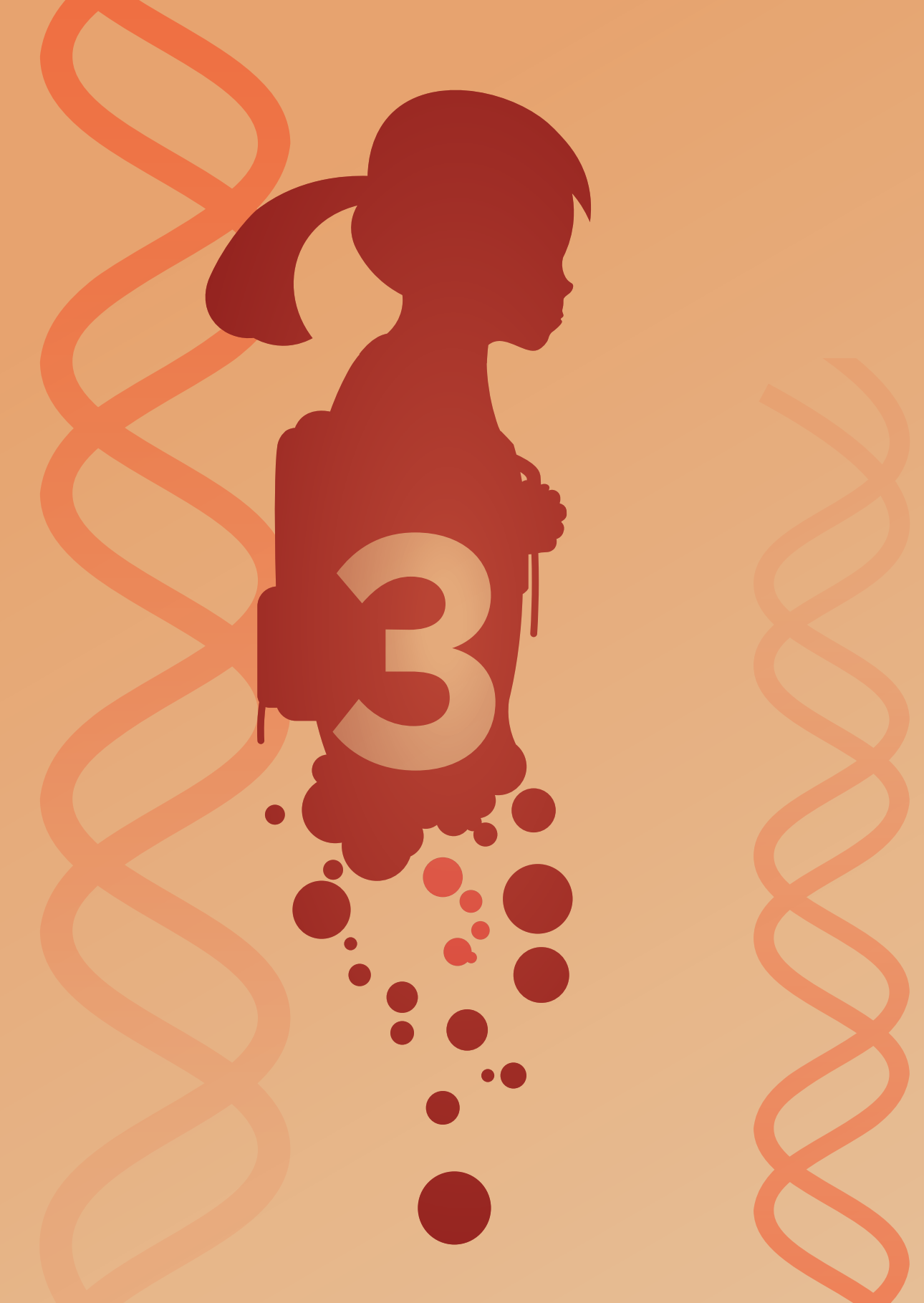
Patient ID	F / M	Primary cancer / treatment protocol	Age Dx1 (y)	Chemotherapy agents (from databases or extracted from protocol)	SCT	RT	Age Dx2/ FU (y)
UPN001	F	Burkitt | LMB 2001	11.5	Cyclophosphamide, ARA-C, MTX, Doxorubicin, Vincristine, Etoposide	No	No	13.2
UPN002	F	B-ALL	3.9	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase	No	No	5.0
UPN003	M	ALL | ALL10, ALL-R3, ALL 11 HR + ADHOC	5.7	At time of FU: MTX, Vincristine, ARA-C, PEG-asparaginase, Daunorubicin, Cyclophosphamide, 6-MP, Doxorubicin, Mitoxantrone, 6-TG | After FU new: ATG, BuFluClo, Teniposide, allogenic MUD-SCT	Yes, 2x	NA	15.7
UPN004	M	B-ALL	15.5	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase, Cyclophosphamide, 6-MP, Doxorubicin	No	No	16.7
UPN005	M	B-ALL	4.3	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase, Cyclophosphamide, 6-MP, Doxorubicin	No	No	5.1
UPN006	F	B-ALL	3.4	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase, Cyclophosphamide, 6-MP	No	No	3.6
UPN007	F	B-ALL	8.3	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase, Cyclophosphamide, 6-MP, Doxorubicin	No	No	9.4
UPN008	F	Osteosarcoma | Euramos1	13.5	Doxorubicin, Cisplatin, MTX | If randomized/ poor response: + Ifosfamide, Etoposide	NA	NA	14.9
UPN009	M	Non low-grade astriocytoma SIOP LGG2004	14.1	None	No	Yes	15.4
UPN010	M	Ewing Ewing2008R3	3.9	Ifosfamide, Doxorubicin, Actinomycin D, Vincristine, Cyclophosphamide, Etoposide	No	Yes	5.7
UPN011	F	Neuroganglioblastoma NBL2009MRG	3.3	Cisplatin, Etoposide, Vindesine, Vincristine, Dacarbazine, Ifosfamide, Doxorubicin, low dose Cyclophosphamide, Retinoic acid	No	Yes	5.9
UPN012	M	Neuroblastoma	4.6	Cisplatin or carboplatin, etoposide, Vindesine, Dacarbazine, Doxorubicin, Ifosfamide, Vincristine, Busulfan, Melfalan.	No	Yes	7.4
UPN013	M	b-thalassemia	NA	Treosulfan, Fludarabine, Thiotepa, ATG, Alemtuzumab	Yes, 2x	NA	5.3
UPN014	M	ALL | ALL11-MRG	6.0	Vincristine, Daunorubicin, Asparaginase, Cyclophosphamide, ARA-C, 6-MP, MTX, Doxorubicin	No	No	7.1
UPN015	F	Lymphoma	NA	NA	NA	NA	15.6
UPN016	F	NHL | ALL VII	NA	Daunorubicin, 6-TG, Vindesine, 6-MP, Asparaginase, Cyclophosphamide, Vincristine, Doxorubicin, Teniposide, MTX, ARA-C, Ifosfamide	No	No	11.2
UPN017	M	AML | ANLL92	5.2	Doxorubicin, Cyclophosphamide, Idarubicin, ARA-C, Vincristine, Mitoxantrone, Etoposide, 6-TG	No	No	15.1
UPN018	M	pre-B ALL | ALL8-MRG	4.6	Vincristine, Daunorubicin, Asparaginase, ARA-C, MTX, 6-MP, Doxorubicin, Cyclophosphamide, 6-TG	No	No	7.8
UPN019	M	AML | ANLL94	7.1	ARA-C, Idarubicin, Etoposide, Mitoxantrone, 6-TG, Vincristine, Doxorubicin, Cyclophosphamide | Conditioning: + Busulfan	Yes	TBI	10.9
UPN020	M	ALL	8.2	NA	NA	NA	14.0

Supplementary Table 1. *continued*

Patient ID	F / M	Primary cancer / treatment protocol	Age Dx1 (y)	Chemotherapy agents (from databases or extracted from protocol)	SCT	RT	Age Dx2/ FU (y)
UPN021		ALL | ALL9-HR	3.9	Vincristine, Daunorubicin, Asparaginase, MTX, 6-MP, Cyclophosphamide, ARA-C	No	NA	5.1
UPN022	F	Ewing sarcoma	NA	NA	NA	NA	9.1
UPN0023	F	T-ALL | ALL10-MRG	9.6	Vincristine, Daunorubicin, Asparaginase, Cyclophosphamide, 6-MP, ARA-C, Leukovorin, MTX, Doxorubicin	No	No	12.2
UPN0024	M	T-LBL | Euro-LB02 III/IV	8.4	Vincristine, Daunorubicin, MTX, Asparaginase, 6-MP, ARA-C, Cyclophosphamide, Doxorubicin, 6-TG,	No	No	10.0

6-TG = thioguanine; ALL = acute lymphoblastic leukaemia; AML = acute myeloid leukaemia; BuFluClo = Busulfan, Fludarabin, Clofarabin; Dx1 = diagnosis 1; Dx2 = diagnosis 2; FU = follow-up; HR = high risk; MUD = matched-unrelated-donor; MRG = medium risk group; MTX = methotrexate; NA = not available; NHL = non-Hodgkin lymphoma; RT = radiotherapy; SCT = stem cell transplantation; TBI = total body irradiation; y = years.

Supplementary Table 2. Per-sample overview of all whole-genome sequenced samples.
Available online (QR code below)


3

Selective pressures of platinum compounds shape the evolution of therapy-related myeloid neoplasms

Eline J.M. Bertrums[1,2,3,*], **Jurrian K. de Kanter**[1,2,*], Lucca L.M. Derks[1,2], Mark Verheul[1,2], Laurianne Trabut[1,2], Markus J. van Roosmalen[1,2], Henrik Hasle[4], Evangelia Antoniou[5,6], Dirk Reinhardt[5,6], Marry M. van den Heuvel-Eibrink[1,7], C. Michel Zwaan[1,3], Bianca F. Goemans[1], 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
[3] Department of Pediatric Oncology/Hematology, Erasmus Medical Center – Sophia Children's Hospital, Rotterdam, the Netherlands
[4] Department of Pediatrics, Aarhus University Hospital, Aarhus, Denmark
[5] Clinic of Pediatrics III, University Hospital of Essen, Essen, Germany
[6] AML-BFM Study Group, Germany
[7] Utrecht University, Utrecht, the Netherlands
* These authors contributed equally

Manuscript under review

Abstract

Therapy-related myeloid neoplasms (t-MN) arise as a complication of chemo- and/or radiotherapy. Although t-MN can occur both in adult and childhood cancer survivors, the mechanisms driving therapy-related leukemogenesis likely vary across different ages. Chemotherapy is thought to induce driver mutations in children, whereas in adults pre-existing mutant clones are selected by the exposure. However, selective pressures induced by chemotherapy early in life are less well studied. Here, we used single-cell whole genome sequencing (WGS) and phylogenetic inference to show that the founding cell of t-MN in children starts expanding after cessation of platinum exposure. In patients with Li-Fraumeni syndrome, characterized by a germline *TP53* mutation, we found that the t-MN already expands during treatment, suggesting that platinum-induced growth inhibition is TP53 dependent. Our results demonstrate that germline aberrations can interact with treatment exposures in inducing t-MN, which is important for the development of more targeted, patient-specific treatment regimens and follow-up.

Introduction

Most chemotherapies act by fatally damaging or inhibiting the synthesis of DNA of cancer cells[1]. However, normal cells are also exposed during treatment, which can promote new carcinogenesis. Therapy-related myeloid neoplasms (t-MN), which include myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML), are hematological disorders that typically occur within 10 years after treatment with cytotoxic therapy for a primary cancer or autoimmune disease[2–4]. Patients with t-MN have a poor prognosis compared to treatment-naïve AML or MDS[5,6], urging the need to develop preventive strategies. T-MN can occur at all ages but remains understudied in children. Nevertheless, it is one of the most prevalent subsequent neoplasms after childhood cancer treatment, besides radiotherapy-induced breast cancers, mostly occurring in females[4]. Previous research that investigated the effects of cytostatic treatment in the hematopoietic system of cancer patients focused on therapy-related clonal hematopoiesis (t-CH) and t-MN in adults[7–11]. Together, these studies propose a model in which chemotherapy exposure mainly induces leukemogenesis by selecting mutated clones that in most cases predated the treatment exposure[7,10,12], and only partly by induction of mutations[12]. Indeed, t-MN driver mutations can already be observed in bone marrow or peripheral blood of adult patients isolated before exposure to cytotoxic therapy[7]. Also, in retrospective studies, specific genes were found to drive t-CH depending on the preceding treatment exposures and which also varied between distinct exposures[10,13]. For example, mutations in DNA damage response (DDR) genes, such as *TP53* and *CHEK2*, were significantly enriched in clones of adults specifically after exposure to cytotoxic therapies, such as platinum drugs and topoisomerase II inhibitors (TOP2i)[10].

Besides selective pressure, DNA damage induced by chemotherapy in healthy cells can increase the genetic diversity in normal tissues during exposure. This can contribute to the development of t-MN by increasing the chance that healthy cells acquire a cancer driving event[9,14,15]. In contrast to adult t-MN, studies suggest that this mechanism occurs often in pediatric t-MN patients, as t-MN driver mutations could not be detected by ultra-deep sequencing methods in samples predating the start of treatment[15,16]. In addition, mutational signature analysis in pediatric t-MN has shown that some driver mutations are directly induced by cytotoxic treatments, such as platinum drugs and thiopurines[14,15,17]. Furthermore, exposure to topoisomerase II inhibitors (TOP2i) is associated with *KMT2A* rearrangements with breakpoints in TOP2-binding regions[18,19]. This association indicates that these rearrangements, that are often found in pediatric t-MN, are likely directly induced by TOP2i. Yet, the evolutionary pressures induced by exposure to these agents and subsequent selection of malignant clones in the blood of children remain unclear.

Here, we aimed to find a model explaining pediatric t-MN development and to subsequently compare this model to the etiology of t-MN in adults. Therefore, we analyzed the genomes of t-MN in 43 children. Using single-cell whole genome sequencing (WGS), we show that although chemotherapy exposure induces the genetic aberrations that drive leukemogenesis, the expansion of the t-MN clone is inhibited by platinum drugs. Subsequently, after finalization of treatment with platinum drugs, the leukemic cell of origin rapidly expands. In Li-Fraumeni Syndrome (LFS) patients, TP53-deficient leukemic clones can expand during platinum drugs exposure, suggesting that platinum-induced growth inhibition is TP53 dependent. Indeed, in LFS patients the t-MN shows a developmental trajectory that is more like that observed in adults.

Results

Pediatric t-MN patient cohort

We included 18 Dutch[14] and 25 German t-MN patient samples that were obtained via a collaboration with the International Berlin-Frankfurt-Münster AML Study Group (I-BFM AML SG). The patients had a variety of first diagnoses and were exposed to different treatment regimens (**Fig. 1a, Table S1**). Most of these children had a primary cancer diagnosis, but four patients received treatment for other underlying diseases (**Table S1**). Except for two patients, all children received chemotherapy (**Fig. 1a**). Although patients UPN009 and IBFM29 were included in this cohort based on clinical diagnosis, their t-MN cannot be specified as chemotherapy-induced, as they received only local radiation or immunosuppressants. Due to the inclusion criteria of our study, 42 patients presented with t-AML and only one with t-MDS. The mean age at t-MN diagnosis was 10.7 years (range 2.4 – 18.4 years). The latency time between first diagnosis and t-MN varied from 0.6 to 11 years (mean 3.4 [95% CI 2.5-4.4], **Fig. 1b, c**). Patients with a hematological malignancy as a first diagnosis had a longer latency time compared to patients with any other primary

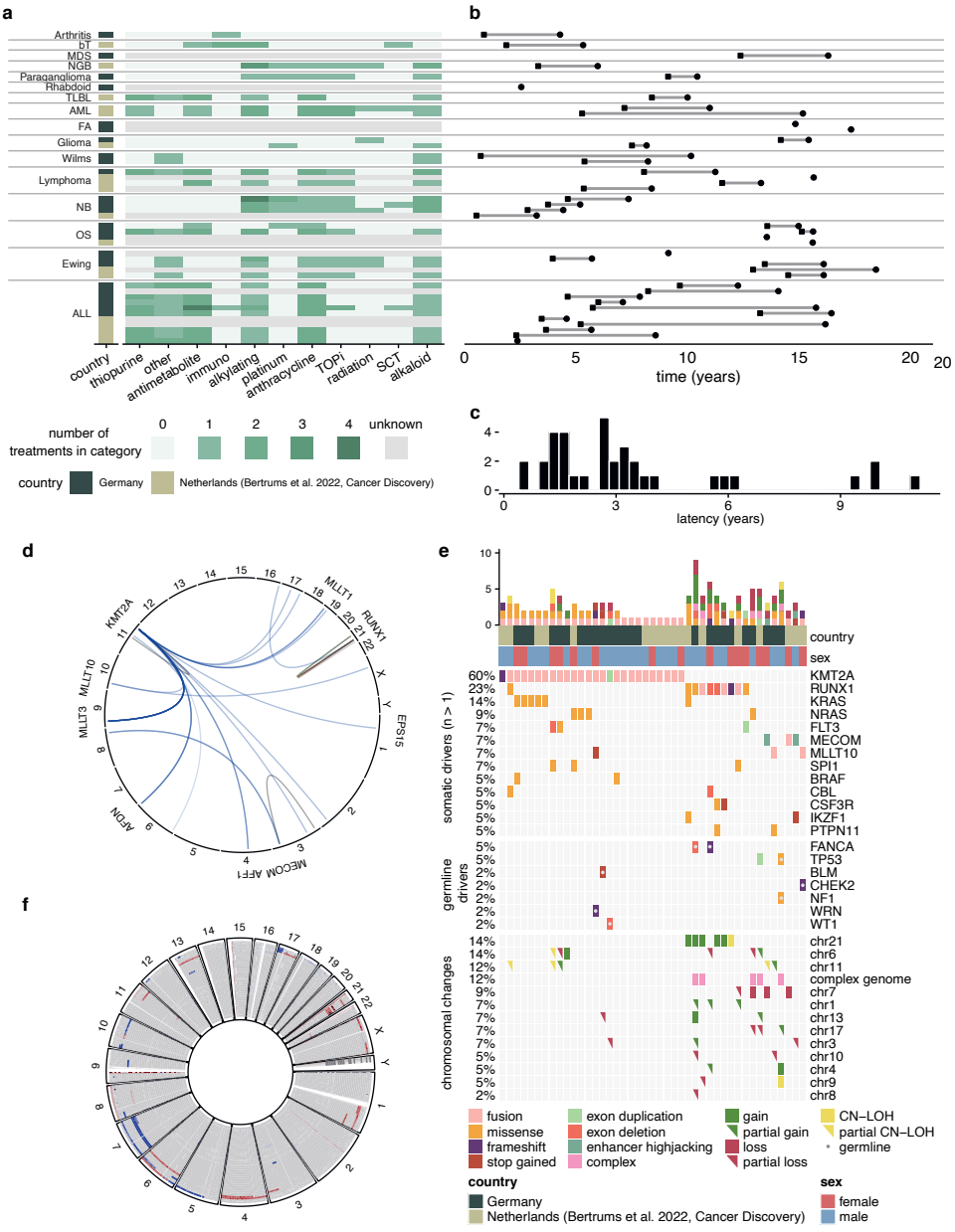


Figure 1. Pediatric t-MN (n=43) is mainly driven by KMT2A fusions.
a) A table depicting the different first diagnoses of t-MN patients and the treatment categories that each patient received. ALL: acute lymphoblastic leukemia; AML: acute myeloid leukemia; bT: beta-thalassemia; FA: Fanconi anemia; MDS: myelodysplastic syndrome; NB: neuroblastoma; NGB: neuroganglioblastoma; OS: osteosarcoma; SCT: allogenic stem cell transplantation; TLBL: T-cell lymphoblastic lymphoma; TOPi: topoisomerase inhibitors. **b)** Per-patient timelines depicting the latency time between the first diagnosis and the t-MN diagnosis. Rows per patient match with (a). **c)** Distribution of latency times in years. **d)** Circos plot of the structural variants (n=70) in t-MN patients that involved at least one cancer gene. *The legend continues on the next page.*

e) Oncoprint depicting the clonal driver events that were present in t-MN samples. The bar plots on top represent the number of driving events present in each sample. Small drivers are only included if they occurred in more than one patient. CN-LOH: copy neutral loss of heterozygosity. **f)** Circos plot depicting the copy number profiles of all t-MN samples.

diagnosis (p = 0.022, **Fig. S1a**). Unfortunately, due to the variety of treatment regimens included in our cohort, this difference in latency time could not be linked to a specific chemotherapeutic compound. We performed WGS on bulk t-MN blasts and used mesenchymal stromal cells (MSCs) or bulk-sorted B-cells of the same patient as a germline control (**Methods, Fig. S2**).

In general, the driver events observed in our cohort were in line with previous studies[14,15] (**Fig. 1d, e, Fig. S1b**). Most recurrent genetic aberrations in the pediatric t-MN samples were structural variants (SVs). Fusions were found in *KMT2A* (58%), *RUNX1* (7%), *MLLT10* (5%) and *MECOM* (2%) (**Fig. 1d, Table S2**). Breakpoints of 9 out of 25 *KMT2A* fusions (36%) overlapped with the 11-bp topoisomerase-associated breakpoint hotspot (**Fig. S1c**)[17,19]. The t-MN with oncogenic fusions usually had fewer chromosomal aberrations than t-MN without fusions (0.6 vs 4.1, p = 3.3*10-5) except for *RUNX1* alterations which co-occurred with chromosome 21 gains in four patients. This co-occurrence has also been described in treatment-naïve AML[20]. Recurrent losses were observed in chromosome 6, 7 and 11 (in 9%, 9% and 7% of the patients respectively, **Fig. 1f**). Compared to adult t-MN, our cohort had a paucity of *TP53* aberrations. A somatic *TP53* mutation was only found in one t-MN (2%), compared to 33% of adult t-MN[7]. Also in de novo pediatric AML, *TP53* aberrations are rare[21].

In the German I-BFM patients, we could investigate germline predisposition variants for which we used a previously published set of childhood cancer-associated genes[22]. Likely pathogenic germline mutations were found in 6/25 t-MN patients (24%), which is higher than previously published in pediatric cancer and pediatric t-MN[15,23], but comparable to adult t-MN[24]. We found a germline aberration in five patients in the genes encoding *BLM*, *WRN*, and *WT1* in one patient each, and *FANCA* in two patients. Patient IBFM22 had a germline aberration in both *TP53* and *NF1*, indicating that this patient had two tumor-predisposition syndromes: LFS and neurofibromatosis type 1 (NF1). In the t-MN blasts, the wild-type allele of *TP53* was lost and the *NF1* mutation was duplicated. In the Dutch cohort we were not able to search for predisposition genes due to ethical constrains, yet the only known germline aberration, based on diagnostic data, was a mutation in *CHEK2*.

Mutational processes underlying t-MN development

A previously established baseline of mutation accumulation during healthy life showed that hematopoietic stem and progenitor cells (HSPCs) normally acquire mutations at a constant rate of 14 to 16 single base substitutions and approximately one indel per year[25,26]. We compared the mutation load of the t-MN blasts to this baseline and

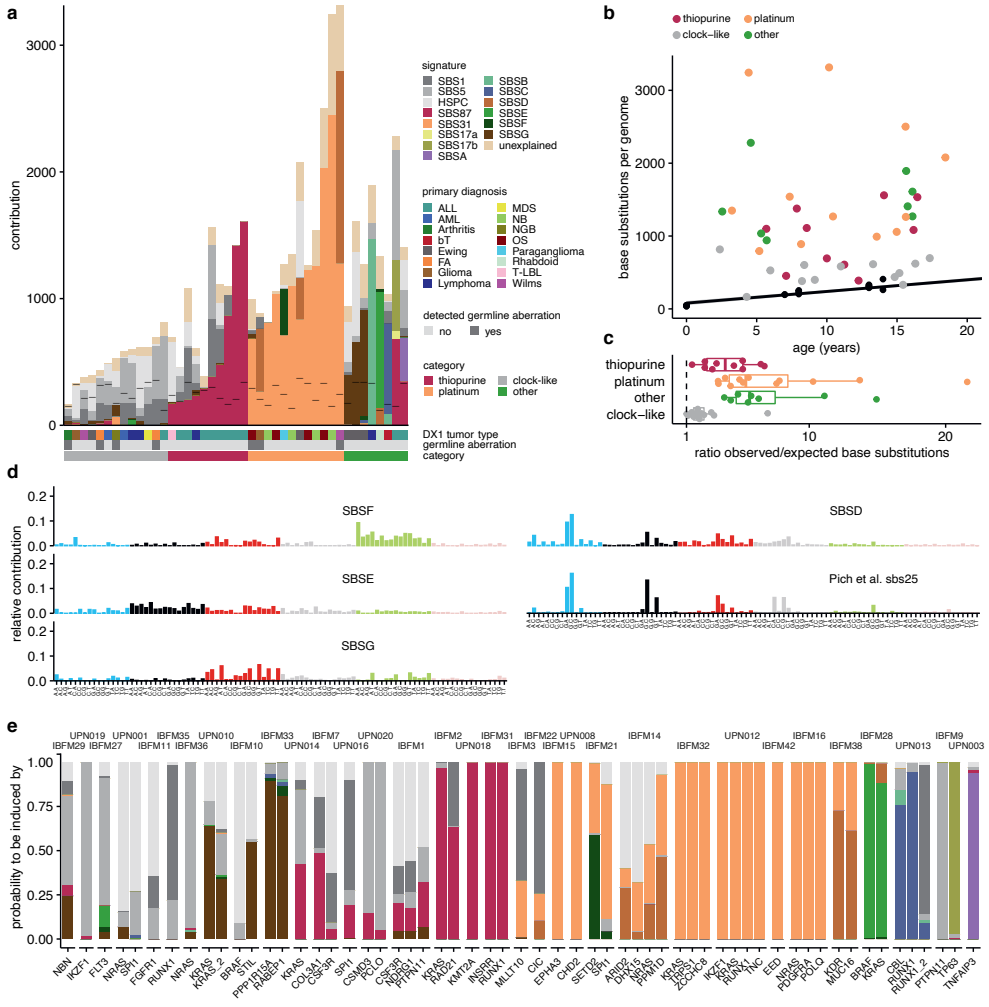


Figure 2. Mutational processes underlying the increased mutation load in pediatric t-MN (n=43)

a) The contribution of each single base substitution signature to t-MN blasts of each patient, obtained after bootstrapped (n=100) refitting of signatures that were extracted by non-negative matrix factorization. The first bar below the plot represents the first diagnosis (abbreviations conform Figure 1a), the second bar notes if a pathogenic germline mutation was found, the third bar represents the treatment category (>150 mutations of that treatment type, or otherwise "clock-like"). **b)** Mutation accumulation of t-MN (colored dots) compared to the baseline of healthy blood cells (black dots). The color is similar to the grouping in (a). **c)** The ratio of the number of observed versus expected single base substitutions in all t-MN samples within a specific signature-category (as in b). **d)** The 96-trinucleotide single base substitution profiles of SBSD-G and the profile of the previously defined signature sbs25 (Pich et al.31). **e)** The probability that different driver mutations (n=42) were caused by treatment-related or clock-like signatures.

found a significant increase (1005 additional mutations, $p<10^{11}$; **Fig. S3A**), similarly to what was previously reported[14].
To elucidate the mutational processes underlying the additional mutations after treatment exposure, we extracted and refitted mutational signatures (**Fig. 2a, Methods**). In 13 out of 43 t-MN cases (30%), all mutations could be exclusively explained by clock-like signatures SBS5 and HSPC, while in the other patients at least 150 mutations could be attributed to an additional signature (**Fig. 2a-c**). We identified the platinum-associated signature SBS31 in platinum-exposed patients[27], and the thiopurine-associated signature SBS87 in thiopurine-exposed patients[28]. In addition, we identified SBS17a/b in the t-MN of one patient (IBFM9) in our cohort, who developed the t-MN after a primary acute lymphoblastic leukemia (ALL). SBS17a/b mutations have previously been attributed to exposure to the drug 5-FU as well as mis-incorporation of oxidized guanines opposite a thymine in the DNA template during replication[29]. Since 5-FU is not used in ALL treatment protocols, the latter process seems more likely.

Furthermore, we identified a signature, here named SBSD, which resembles a carboplatin-associated signature previously described by Pich et al. in metastases of adult solid tumors[30] (sbs25, cosine similarity 0.87, **Fig. 2d**). The 96-trinucleotide profile has similarities to SBS31 and SBS35, which are both platinum-induced signatures. SBS31 and SBSD both co-occur with a platinum-induced double base substitution signature DBS5 in our dataset (**Fig. S3b-d**). SBSD was mainly present in four patients (IBFM22, IBFM38, IBFM14 and IBFM28). Interestingly, IBFM22, who was treated with carboplatin for *NF1*-related opticus glioma, also harbored a germline *TP53* mutation, characteristic of LFS. TP53 is essential in the G1/S cell cycle checkpoint and is activated upon a variety of cellular stresses, including DNA damage[31]. *TP53* is the most commonly mutated gene in human cancer[32] and mutations have been associated with platinum-resistance[33], potentially explaining the distinct signature. Interestingly, the t-MN of IBFM14 had a heterozygous loss of chromosome 17p that harbors the *TP53* gene and IBFM38 had a germline *WT1* mutation, which has been described to impact downstream factors of TP53[34], and thus likely increases resistance to DNA damaging agents. Like IBFM22, IBFM38 (Wilms tumor) and IBFM28 (atypical rhabdoid tumor) were highly likely treated with carboplatin, according to the applicable treatment protocols at the time the patient was treated, whereas this was unclear for IBFM14 (Ewing sarcoma). Notably, also patient IBFM21 ($TP53^{+/+}$) received carboplatin therapy, and this patient did not show SBSD-related mutations. These findings suggest that the type of carboplatin-induced mutations might be influenced by TP53 function.

To further validate this, we compared $TP53^{+/+}$, $TP53^{+/-}$ and $TP53^{-/-}$ metastases of cancers that had been treated with carboplatin or cisplatin from a cohort of 4,853 metastases from 4,711 patients previously described by Priestley et al.[35]. We checked the contribution of the SBS31, SBS35 and the carboplatin-associated sbs25 that was

found by Pich et al. and is similar to SBSD in this cohort. After carboplatin exposure, fewer *TP53*$^{-/-}$ and *TP53*$^{-/+}$ tumors contained SBS35 mutations compared to *TP53*$^{+/+}$ tumors (22%, 44% and 55% respectively, not significant). In contrast, more *TP53*$^{-/-}$ and *TP53*$^{-/+}$ tumors had sbs25 (SBSD) mutations than TP53+/+ tumors (55%, 38% and 28%, respectively). After cisplatin exposure SBS31 was the dominant signature, independent of *TP53* status (**Fig. S4**). Although not confirmative, these results are in line with the idea that *TP53* status influences the type of mutations that accumulate during carboplatin exposure.

Finally, we identified three novel signatures with unique 96-trinucleotide profiles, which we here named SBSE, SBSF and SBSG (**Fig. 2d**). Interestingly, SBSG was identified in three t-MN that developed after Ewing sarcoma, suggesting a potential association with the treatment regimen. As SBSE and SBSF only occurred in the t-MN of one patient each (IBFM28 and IBFM21, respectively), it remains challenging to elucidate the underlying process. While the profile of SBSE shows mainly C>G changes, in SBSF mainly T>C changes are seen, which is in line with exposure to alkylating agents[36].

We used a previously established method to calculate the probability that the t-MN driving mutations were caused by each detected signature[37,38]. This analysis revealed that clock-like signatures could explain 29% of the single nucleotide variant drivers. On the other hand, 45% of the driver mutations could be attributed to a non-clock-like signature with >70% likelihood (**Fig. 2e**), showing that also in our cohort treatments did induce small drivers.

Clonal evolution of t-MN under platinum exposure

To study the selective pressure of platinum compounds in the blood of t-MN patients, we performed retrospective lineage tracing in multiple patients. WGS on bulk, single t-MN blasts, and single HSPCs was performed on DNA that was extracted from clonally expanded cells or directly amplified from single cells by primary template-directed amplification (PTA)[39]. By comparing shared and individual mutations in all sequenced cells from the same patient, we could construct phylogenetic trees in which each split of a branch represents a cell division (**Fig. 3**). These trees can be used to study the evolution of the t-MN over time, by timing the moment of the t-MN expansion and by determining the mutational processes that were active before, during and after this expansion. Clonal mutations were those shared between all the t-MN blasts. These accumulated before expansion of the initial leukemic cell. Subclonal mutations were those shared between a subset of single t-MN blasts and accumulated during expansion of the leukemic clone. Finally, private mutations were unique to single blasts and accumulated most recently during t-MN expansion. Of note, in four patients some of the immunophenotypically HSPC-like cells shared all the t-MN drivers and other clonal mutations with the bulk t-MN blasts and were thus in the phylogenetic analysis considered t-MN blasts.

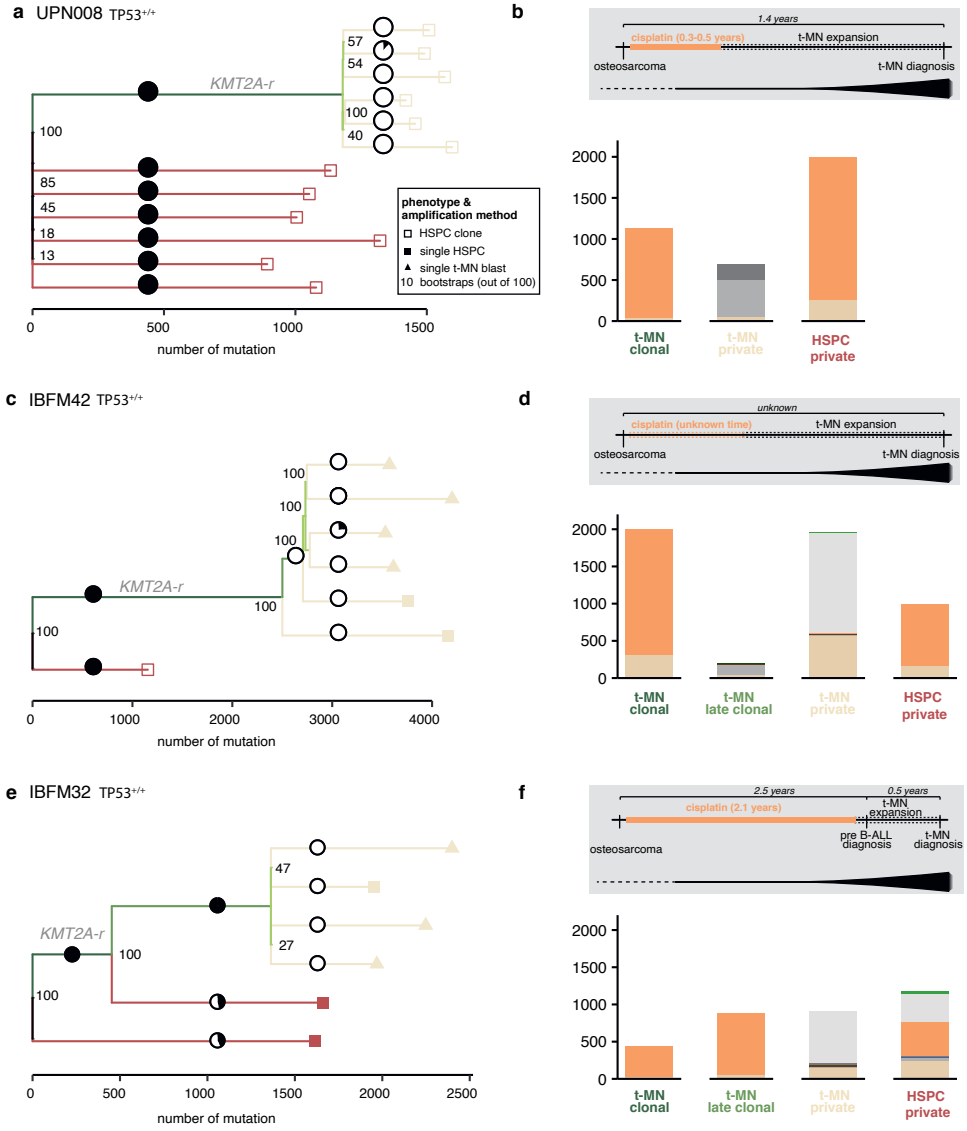


Figure 3. Clonal evolution of t-MN under platinum treatment in patients without germline TP53 aberrations.

a) Phylogenetic tree of clonally expanded HSPCs and bulk t-MN blasts of patient UPN008 (*TP53* wild-type). Pie charts indicate the contribution of SBS31 after strict refitting (max_delta < 0.01). The colors of the branches correspond to the type of mutation, clonal, subclonal, or private, which is annotated in the same color text. **b)** Signature contribution of the mutations in the corresponding branches in the lineage tree on the left. The private HSPC mutations were subsampled to 2000 mutations for visual purposes. Top: schematic overview of the timeline of the different diagnoses and treatment, including the timing of t-MN development. **c)** simar to (a), but for patient IBFM42 (TP53 wild-type). Also single cell sequenced HSPCs (black squares) and t-MN blasts (black triangles) are included. Dotted lines represent sequenced cells in which the 5x coverage was less than 85% of the genome.

The legend continues on the next page

d) Similar to (b), but for patient IBFM42. **e)** similar to (a), but for patient IBFM32. The black diamond indicates the clonal mutations in the t-MN bulk blasts that were not found in any of the sequenced single blasts. The AML blast population of this patient showed two distinct immunophenotypes (CD34+ and CD34-). The single AML blasts that were sequenced were all CD34+. The KRAS mutation was not present in any of these blasts, but its presence could be confirmed in DNA of CD34- blasts with PCR and Sanger sequencing. **f)** Similar to (b), but for patient IBFM32.

We ran Cellphy[40] on previously published data of UPN008[14], a patient who developed a t-MN after platinum-treatment (n=12 HSPCs). The bulk t-MN blasts of this patient harbored many clonal platinum-related mutations, only 11 subclonal mutations, and no platinum-related private mutations (**Fig. 3a, b**). This observation suggested that the t-MN clone started expanding after the end of platinum treatment. Based on the length of the individual branches and the minimum latency of 1.1 years between the end of platinum treatment and t-MN diagnosis, the mutation rate in the blasts was at least 297 sbs/year. The 11 subclonal mutations would suggests that the detected cell divisions happened within 14 days. Of note, it is unknown if the mutation rate in t-MN blasts is constant, it is likely that the division happened within a relatively short period of time. The phylogenies of patients IBFM32 (n=3 HSPCs, n=3 blasts) and IBFM42 (n=3 HSPCs, n=4 blasts), both treated with cisplatin, showed a similar pattern. For these patients, besides HSPCs, we also included single blasts in our analysis. Apart from a single division detected in the middle of the t-MN evolution of IBFM32, both patients had long clonal branches and short subclonal branches (**Fig. 3c-f**). Similar as in patient UPN008, all clonal, but only few subclonal and private mutations were platinum-related, suggesting that the expansion of the t-MN clones in both patients happened after the end of platinum treatment. These results support the idea that platinum exposure inhibits the expansion of the initial leukemic clone, and that this expansion thus starts when the exposure to platinum ends.

The role of TP53 in clonal evolution under the selective pressure of platinum exposure
As discussed above, we found that TP53-deficient cells that were exposed to carboplatin accumulated a different mutational signature (SBSD) compared to TP53-proficient cells (SBS31). In addition, previous literature indicates a clonal advantage of *TP53*-mutated cells under platinum treatment[33]. Therefore, we investigated the clonal evolution of t-MN in the carboplatin-treated LFS patient IBFM22 and compared this to *TP53* wild-type platinum-treated t-MN patients.

In the phylogeny of patient IBFM22 (n=5 HSPCs, n=5 blasts), fewer clonal mutations were present. In the clonal branch, also chromosomal copy number changes were present, among which a 17p deletion that resulted in the loss of the *TP53* wild-type allele. In addition, only few of these trunk mutations could be attributed to platinum treatment, indicating that the t-MN clone started expanding very early during treatment (**Fig 4a, b**). In contrast to the $TP53^{+/+}$ t-MN cases, two subclonal branches contained hundreds of mutations, all of which were platinum-related. Furthermore, we could observe multiple branching points that occurred during platinum exposure,

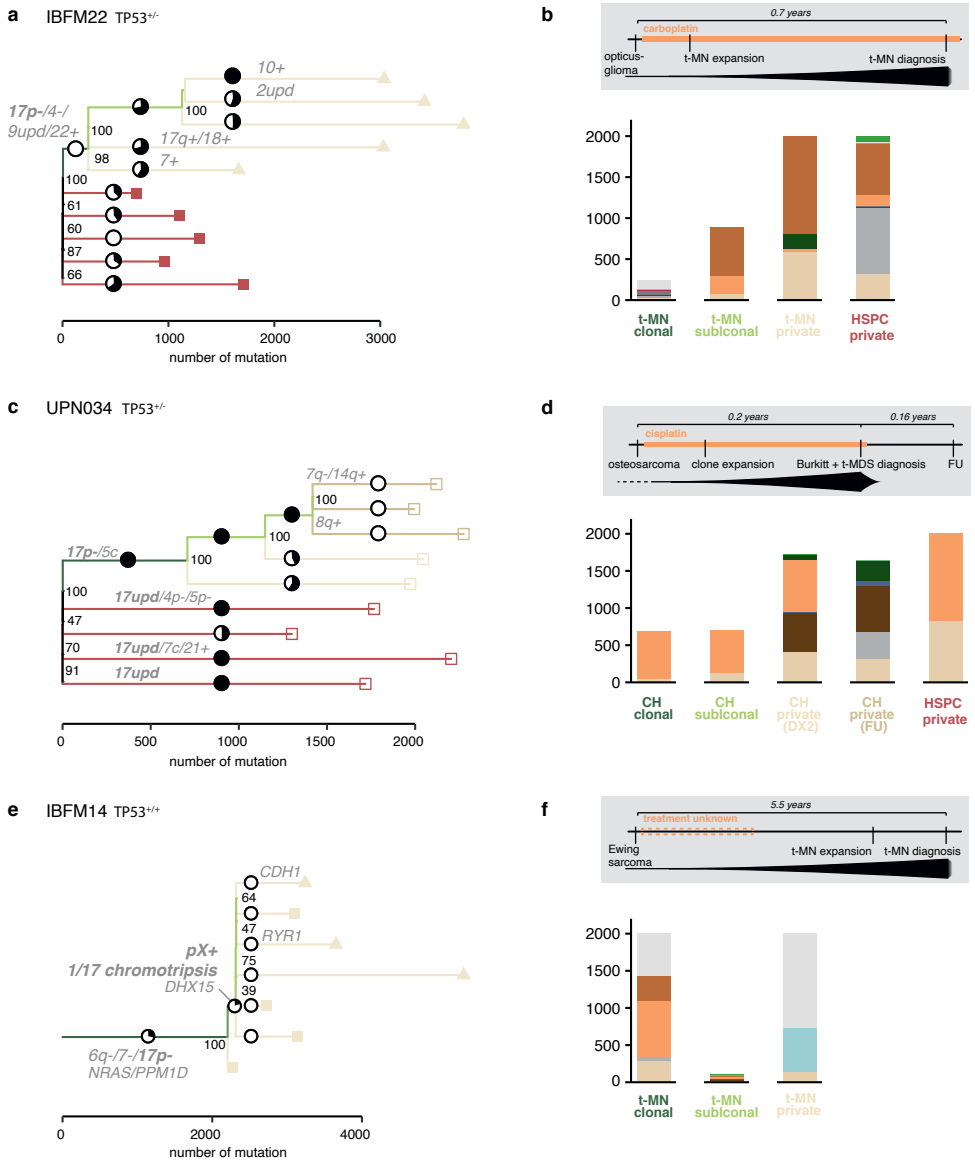


Figure 4. Clonal evolution of t-MN under platinum treatment in patients with germline *TP53* aberrations.

a) Phylogenetic tree of single HSPCs (black squares) and bulk and single (black triangles) t-MN blasts of patient IBFM22. Pie charts indicate the contribution of SBSD after strict refitting (max_delta < 0.01). Dotted lines represent sequenced cells in which the 5x coverage was less than 85% of the genome. **b)** Signature contribution of the mutations in the corresponding branches in the lineage tree on the left. Top: schematic overview of the timeline of the different diagnoses and treatment, including the timing of t-MN development. **c)** simar to (a), but for patient UPN034. Clonally expanded HSPCs (white squares) were sequenced. **d)** Similar to (b), but for patient UPN034. The HSPC private mutations were subsampled to 2000 mutations for visual purposes. CH: clonal hematopoiesis. **e)** Similar to (a), but for patient IBFM14. **f)** Similar to (b), but for patient IBFM14.

as 60% of private mutations harbored platinum-related mutations, indicating that the expansion of the t-MN in this case occurred during treatment. This model was further strengthened by the short latency of 0.7 years between first diagnosis and t-MN development in patient IBFM22, which was still within the carboplatin treatment window. This latency time was longer for patients UPN008 (1.4 years) and IBFM32 (3 years), and unknown for patient IBFM42. In conclusion, in *TP53*$^{-/-}$ t-MN the timing and speed of the clonal expansion appeared to be different compared to *TP53*$^{+/+}$ t-MN.

To confirm the proposed interaction between TP53 deficiency and the evolutionary pressures that are induced by platinum treatment, we studied the post-treatment bone marrow sample of another LFS patient (UPN034, n=9 HSPCs), who had been previously treated for osteosarcoma (DX1) with, among other treatments, cisplatin. Less than three months later, during the treatment for osteosarcoma, this patient was diagnosed simultaneously with Burkitt lymphoma and t-MDS with less than 5% leukemic blasts in the bone marrow (DX2). We sequenced nine clonally expanded HSPCs, six from the time of Burkitt, and three from follow-up (FU) two months later. Interestingly, five HSPC clones, including all three FU clones, shared the same event that resulted in the loss of the wild-type TP53 allele (**Fig. S5a**). This is suggestive for the presence of CH at time of DX2 in which one HSPC expanded at a disproportionately higher rate than other HSPCs[41]. Comparable to IBFM22, the subclonal branches were hundreds of mutations long and were platinum related (SBS31), suggesting that also this clone expanded during treatment (**Fig. 4c, d**). Interestingly, all cells at time of FU arose from the same clone and the private mutations in these cells occurred after the end of platinum treatment.

Since we were not able to perform WGS on the bulk t-MN blasts of UPN034 in our study, we requested single nucleotide polymorphism (SNP) and karyotype assays data from our diagnostics department. This way, we could determine how the HSPCs were related to the t-MN. One sequenced HSPC shared a RUNX1 aberration and 7q loss with the t-MN, but not the additional 20q loss. These findings imply that this HSPC is a pre-leukemic cell. Notably, the *TP53* wild-type allele was lost in the HSPCs by four independent 17p loss events which all had unique breakpoints. This convergent evolution of multiple clones that independently lost the wild-type allele of *TP53* indicates a strong selective pressure to lose TP53 function in this patient[42] (**Fig. S5a**). To investigate whether this selective pressure was due to the treatment or already existed before, we investigated WGS data of bulk peripheral blood at the time of primary osteosarcoma diagnosis. Copy number variant (CNV) analysis of these data revealed that a variety of 17p copy-number neutral loss of heterozygosity (CN-LOH) events were already present in the blood before the start of any treatment (at time of DX1). This resulted in the loss of a copy of *TP53* in 38% of the blood cells, indicating that TP53 wild-type allele loss events were already present before treatment (**Fig. S5b**).

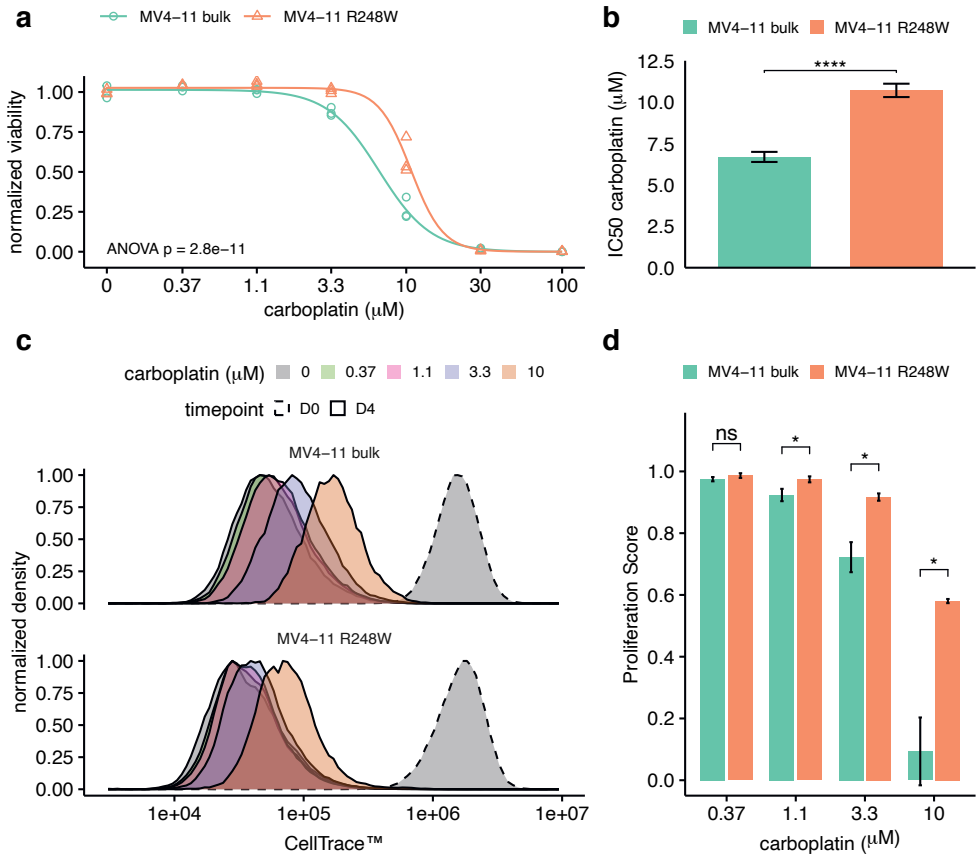


Figure 5. TP53 deficiency enables increased proliferation under platinum treatment.
a) Dose-response curves of MV4-11bulk (circles, each the mean of 3 technical replicates) and MV4-11^{R248W} (triangles, each the mean of 3 technical replicates), based on the DAPI-negative fraction of single cells, n = 3 biological replicates per cell line. The complete dose-response model was tested against the null model, lacking cell line information (ANOVA). **b)** The IC50 values of carboplatin treatment per line, extracted from the dose-response models depicted in (a). The comparison of the IC50 values is based on a z-test and error-bars represent the standard error. **c)** CellTrace signal per treatment condition, normalized to unit area, for MV4-11^{bulk} cells (top) and MV4-11^{R428W} cells (bottom). **d)** Proliferation Score per treatment condition for MV4-11^{bulk} and MV4-11R^{248W}. The scores of each cell line were compared within treatment conditions using a holm's corrected one-sided T-test. Error bars represent standard deviation of the mean of three independent experiments. **** p<0.0001, * p<0.05.

Subsequently, we investigated if a clone could also escape platinum-induced inhibition when losing only one allele of *TP53*. Therefore, we performed WGS on single HSPCs (n=4) and t-MN blasts (n=3) of patient IBFM14, whose t-MN had a heterozygous loss of chromosome 17p, involving *TP53*, and no additional *TP53* mutation. Although the patient history is unclear about platinum treatment, mutational signature analysis of the t-MN revealed contribution of both SBS31 and SBSD, indicative of carboplatin exposure. Notably, the private mutations of the single

t-MN blasts had contribution of SBS2, indicative of APOBEC activity, which is an uncommon signature in pediatric cancer, including AML[43,44]. Other than that, the t-MN in this patient followed the same pattern as the three *TP53*$^{+/+}$ t-MN, with many platinum-related clonal mutations, few subclonal mutation, and no platinum-related private mutations, indicative of expansion after the end of platinum treatment (**Fig. 4e, f**). This confirms that only t-MN clones that have no *TP53* wild-type allele can escape platinum-induced inhibition.

Finally, we validated the effect of TP53 deficiency on cell proliferation under platinum treatment using MV4-11^{bulk}, a pediatric acute monocytic leukemia cell line with a subclonal *TP53* R248W mutation (36% of the total alleles), and MV4-11^{R248W} in which the R248W variant was homozygous in all cells[45,46]. In line with our *in vivo* findings, MV4-11^{R248W} was more resistant to carboplatin treatment than MV4-11^{bulk} (IC50 10.7μM vs. 6.7μM, p<10-10, **Fig. 5a, b, Fig. S6**). Next, we used a single pulse of CellTrace dye, which is equally distributed over daughter cells during cell divisions, to track proliferation during *in vitro* treatment. From the fluorescent intensity after four days of treatment, we calculated a Proliferation Score (**Methods**). MV4-11^{R248W} showed a dose-dependent increase in proliferation compared to MV4-11^{bulk}, confirming that TP53-deficiency leads to platinum resistance by enhancing proliferation during treatment (**Fig. 5c, d**).

Discussion

Previous studies have described differences in the evolution of t-MN in children and adults[7,14,15]. Whereas in adults the founding cell is often already present before treatment, and subsequently selected during chemotherapy, in children the driving event of the t-MN is likely induced during therapy, and the t-MN blasts probably expand afterwards[7,14]. Here, we show that platinum-based treatment inhibits the clonal expansion of the t-MN blasts. The evolution of *TP53*$^{+/+}$ shows expansion of a single clone likely in a short period of time, after overcoming a selective pressure, in this case by cessation of platinum-based treatment. On the other hand, pediatric *TP53*$^{-/-}$ t-MN are able to expand during platinum exposure and therefore have an evolutionary trajectory that is more similar to those described in adult t-MN[7] and different from pediatric *TP5*$^{3+/+}$ t-MN. As platinum drugs are usually administered in short intervals during the treatment protocol, it is unclear whether *TP53*-$^{/-}$ cells expand *in vivo* during exposure itself, similar to the MV4-11 cells *in vitro*, or whether they recover faster and expand during treatment intervals. Overall, more *TP53*$^{-\backslash-}$ t-MN cells are able to survive during treatment. The observations that a complete loss of *TP53* is needed for this selective advantage would indicate that also in adult *TP53*-aberrant t-MN the wild-type *TP53* copy is already lost before treatment, opposite to a previous model by Wong et al.[7]. This hypothesis is supported by the events that caused loss of wild-type *TP53* in UPN034 before treatment.

Following our findings, a model arises in which comparable selective pressures are present during platinum-based treatment in adults and children. However, due to a higher number of mutations that accumulated due to aging and potentially environmental mutagenic exposures, adults would have a higher chance to already have *TP53*-mutated cells in their blood before the start of treatment[25,47,48]. The chance would thus be relatively high in adults that one of these cells loses the *TP53* wild-type allele before treatment, and therefore gains a competitive advantage and develops into t-MN. In contrast, in children *TP53* aberrations are hardly present in blood. Thus, only cells in which both *TP53* wild-type alleles are lost prior to or in the beginning of treatment, can then expand during treatment. As this chance is much lower in children, a smaller part of pediatric compared to adult t-MN harbor *TP53* aberrations. On the other hand, in children with germline *TP53* aberrations this chance is higher, thus the clonal evolution of the t-MN can mimic that of adults. Hence, in these children possibly even more t-MN are driven by TP53 deficiency and the loss of the *TP53* wild-type allele, as in theory every HSPC in their blood could lose the wild-type *TP53* allele. Finally, for cells with other, non-*TP53*, aberrations, such as *KMT2A* fusions, expansion mostly occurs after the end of platinum-based treatment. When they do so, the expansion seems to happen rather quickly, conceivably because the niche is very sparsely populated.

In addition to a different clonal trajectory, *TP53*$^{-/-}$ tumors accumulate distinct mutational profiles under carboplatin treatment. This signature (SBSD/sbs25) was previously identified in the metastases of treated cancer patients and linked to carboplatin exposure[35]. Here, we confirm this relationship and show enrichment of SBSD in both t-MN and metastases with TP53 aberrations. It is very likely that more interactions between germline aberrations and treatment exposure are present. Due to the relative rarity of germline aberrations, and the large variety of different treatments that are used, large datasets are needed for the identification of these interactions. Interactions between germline mutations and treatments, such as the TP53 – carboplatin interaction that we reveal here, could have clinical implications, not only for future choices of treatment regimens, but also for follow-up of these patients after treatment.

Methods

Patient samples

Patient samples were collected via the biobank of the Princess Máxima Center for Pediatric oncology, with ethical approval under proposal PMCLAB2020.151 and via a collaboration with the I-BFM AML SG from the German AML-BFM study group in accordance with the Declaration of Helsinki. Informed consents were obtained from all participants.

Sample work-up

Bone marrow mononuclear cells were stained for fluorescence-activated cell

sorting (FACS) after thawing. Hematopoietic stem and progenitor cells (HSPCs) were identified using the following surface markers: $Lin^-CD11c^-CD16^-CD34^+$, $CD38^-$/CD45RA ("HSPC mix"). t-MN blasts were defined based on diagnostic immunophenotype data if available. If immunophenotype data were unavailable, cells were stained with the same antibody panel including CD33, CD34 and CD38. B-and T-cells were identified using the following surface markers: CD3, CD4, CD20, CD33, CD34 ('mature mix').

Blasts and HSPCs were purified on a SH800S Cell Sorter (Sony). First, blasts were sorted in bulk for DNA isolation after which single HSPCs and t-MN blasts were index sorted in a 96-well plate prepared with PTA-buffer. Additionally, single HSPCs were sorted in a 384-well plate prepared with 75 µL HSPC culture medium per well. HSPC culture medium consisted of StemSpan SFEM medium (Stemcell technologies) supplemented with SCF (100 ng/mL), Flt3-ligand (100 ng/mL), IL-6 (20 ng/mL), IL-3 (10 ng/mL), TPO (50 ng/mL), UM729 (0.5µM) and Stemregenin (750nM).

HSPCs sorted in HSPC culture medium were cultured for 4-7 weeks at 37°C, 5% CO2 before harvesting. Mesenchymal stromal cells (MSCs) were cultured from a bone marrow fraction of 500,000 cells/well in 12-well culture dishes with DMEM-F12 medium (GIBCO), supplemented with 10% FCS. Medium was refreshed every other day to remove non-adherent cells and MSCs were harvested when confluent, after approximately 2-3 weeks.

FACS antibodies

All antibodies were obtained from Biolegend, except for CD13 (Biosciences). Antibodies used for t-MN blast and HSPC populations: CD34-BV421 (clone 561, 1:20), lineage (CD3/CD14/CD19/CD20/CD56)-FITC (clones UCHT1, HCD14, HIB19, 2H7, HCD56, 1:20), CD38-PE (clone HIT2, 1:50), CD90-APC (clone 5E10, 1:200), CD45RA-PerCP/Cy5.5 (clone HI100, 1:20), CD33-PE/Cy7 (clone WM53, 1:100), CD49f-PE/Cy7 (clone GoH3, 1:100), CD16-FITC (clone 3G8, 1:100), CD11c-FITC (clone 3.9, 1:20), CD123-Pe/Cy7 (clone 6H6, 1:100), CD13-PerCP/Cy5.5 (Biosciences, clone WM15, 1:20), CD14-APC (clone HCD14).

For the B- and T-cell sort the following antibodies, obtained from Biolegend, were used: CD3-PE/Cy7 (clone SK7), CD4-PerCP/Cy5.5 (clone OKT), CD20-BV421 (clone 2H7), CD33-APC (clone WM53), CD34-APC (clone 561).

DNA isolation and WGS

DNA was isolated from cell pellets of blasts, MSCs and clonally expanded HSPCs ("HSPC clones") using the DNeasy DNA Micro Kit (Quagen), following the manufacturer's instructions. The standard protocol was slightly adjusted by adding 2µL RNase A (Qiagen) during the lysis step and eluting DNA in 50µL low EDTA TE

buffer (10mM Tris, 0.1mM EDTA, G Biosciences).

DNA from single HSPCs and blasts was amplified using the ResolveDNA™ WholeGenome Amplification Kit (BioSkryb) according to the manufacturer's instructions.

For each sample, DNA libraries for Illumina sequencing were generated from at least 45 ng genomic DNA using standard protocols. For PTA-amplified DNA at least 500ng genomic DNA was used. The libraries were sequenced on Novaseq 6000 sequencers (2x150bp) at a depth of 15x (clones/single cells) or 30x (bulk t-MN and control samples). Two t-MN bulk blast DNA pallets (15 and 22% purity) were sequenced at a depth of 90x.

Sample combinations during processing

For the characterization of the genomic landscape of the t-MN, for all patients the subsequent steps were performed on the bulk t-MN together with the matched normal sample for each patient. For patients for which single cells were sequenced the mutation calling was performed again on the single cells, HSPC clones, bulk t-MN, and matched normal bulk per patient.

Read mapping

Sequencing reads were first mapped to genome GRCh38 using Burrows-Wheeler Aligner (bwa) v0.7.17 using "bwa mem –M –c100", then duplicates were marked using Sambamba v0.6.8 and base recalibration was performed using GATK's BaseRecalibrationTable and BaseRecalibration. All GATK tools were from version 4.1.3.0.

Mutation calling, filtering, and annotation

Mutation calling was performed with GATK's HapoTypeCaller on each set of samples. Mutation filtering was performed using GATK's SelectVariant with options "--select-type SNP --select-type NO_VARIATION --select-type INDEL --select-type MIXED'" on the resulting multi-sample VCFs. Next, VariantFiltration was run with the following options: –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" -cluster 3 -window 10." And finally –filter-expression "QD < 2.0". For 90x samples the last expression was replaced by –filter-expression "QD < 1.0", to prevent somatic mutations with a high coverage to be filtered out. The mutations with a QD between 1 and 2 had a similar mutational pattern compared to the mutations with a QD value higher than 2, and were manually inspected, confirming that they were not artifacts. Annotation of variants was done with SNPEffFilter, SNPSiftDbnsfp (dbNSFP3.2a),

GATK VariantAnnotator (COSMIC v.89) and SNPSiftAnnotate (GoNL release 5). A full pipeline description is available at www.github.com/UMCUGenetics/NF-IAP.

Somatic mutation filtering

SMuRF v3.0.0 was used for obtaining high-confidence clonal somatic mutation calls (www.github.com/ToolsVanBox/SmuRF). These were mutations that (A) are positioned on autosomal chromosomes, (B) have a GATK phred-scaled quality score R 100, (C) have a mapping quality of 60 (30x coverage) or 55 or higher (15x), (D) had a base coverage of at least 10 (30x) or 5 (15x), (E) had a GATK genotype quality of 99 (heterozygous) or 10 (homozygous) in both the sample and paired control (if available), (F) were clonal, and thus had a variant allele frequency (VAF) of > 0.3 (30x), 0.15 (15x) or 0.07 (90x), (G) had no evidence in the paired control sample if available.

Mutation filtering for PTA single-cell WGS data

For single cells, our in-house developed pipeline PTATO was applied, which uses SMuRF v3.0.0 as well as germline mutations combined with a random forest and walker to separate real somatic mutations from amplification-induced artifacts. For a full description, see our recent publication[49]. The final set of PTATO mutations was only used to filter the single-sample end branches of the phylogenetic trees (see below, section "Phylogenetic tree construction").

Mutation filtering for bulk t-MN without a paired normal

For bulk t-MN samples without a paired normal (n = 2), or with evidence of blast contamination in the normal (n = 2), all HSPC clones were used for filtering as described previously14. In these cases, a mutation identified by SMuRF to be in the bulk t-MN was excluded if it (a) was clonally present in all samples that passed QC for that mutation, (b) was subclonally present in any sample, or (c) was not confidently absent in at least one sample.

Baseline and t-MN mutation load normalization

Somatic mutations were re-called using SMuRF v3.0.0 in healthy HSPC samples from a previously published baseline. The number of autosomal mutations in the baseline and t-MN samples were corrected based on GATK's CallableLoci's CALLABLE length. A linear mixed-effects model was fit on the baseline samples and the slope and intercept of this line were used to calculate the expected mutation load for each t-MN sample.

Mutational signature extraction and refitting

MutationalPatterns package v3.6.0 was used for mutational signature extraction and refitting. As signature extraction becomes more robust when more samples are included, we combined the SBS of all t-MN bulk samples with previously published data of 34 healthy HSPCs of healthy individuals of different ages, which were also used

for the baseline (see above)[25]. Extraction was done by applying the extract_signatures function with options "rank = 12, nrun = 100". Then, the signatures that correlated to signatures from the COSMIC database v3.2 and previously identified signatures in the Dutch part of this cohort[14], with a cosine similarity of 0.8 or higher were substituted by the highest correlating signature. Signatures that could be reconstructed by three or fewer COSMIC signatures (cosine >= 0.85) were substituted by those known signatures. One of the signatures was separated in SBS87, SBS17a and SBS17b.

Refitting was done with the extracted signatures using fit_to_signatures_bootstrapped with options "n_boots = 100, max_delta = 0.05". The contributions from the 100 refits were averaged and the difference between the total numbers of mutations in each sample and the sum of refitted mutations was categorized as "unexplained".

Small driver events

Single base substitution and indel driver events were extracted from the output of SMuRF v3.0.0. Only exonic mutations with MODERATE or HIGH impact according to the SnpEff annotation were considered. In addition, mutations in genes from the COSMIC cancer gene consensus v97 and pediatric AML drivers genes were included. Finally, missense, nonsense, frameshift, insertions and deletions were considered as driver events. Driver events (including SVs and CNVs) were visualized using the R package ComplexHeatmap v.2.12.0[50].

Structural variation calling

The Hartwig Medical Foundation's gridss-purple-linx pipeline v1.3.2 was applied on the bulk t-MN blast samples and their paired normal to call somatic structural variants (SVs) and determine copy number alterations (CAN) with options '—amber_tumour_only "true" –cobalt _tumour_only "true" –purple_tumour_only "true"'. All structural variants were checked in IGV and false positives were filtered out. The sub-packages "amber" and "cobalt" that are part of this pipeline were also applied on the bulk blood WGS from time of primary diagnosis of patient UPN034, which was obtained from the diagnostics department.

Germline SVs were extracted from unfiltered GRIDSS VCFs using bcftools filter v1.14[51] in multiple steps using the following filters. 1) -i 'FILTER=="PASS" && BMQ > 40 && FORMAT/RP > 10 && MQ>55 && INFO/ASQ > 0', 2) -i 'INFO/ASSR > 20 | INFO/ASRP > 20' $QUAL > $QUAL2. These variants passing these filters were then overlapped with driver genes from a pediatric cancer WGS dataset from Grobner et al.[22] using "bedtools[52] intersect -header -wa". Finally, the final set of SVs was manually investigated in IGV and only SVs with supporting reads in the matched control and t-MN sample and those overlapping with at least one exon of a cancer gene were reported. Structural variants were visualized using the R package "circlize" v0.4.15. Breakpoints were visualized using the R package "Gviz" 1.40.1[53].

Phylogenetic tree construction

Phylogenetic reconstruction was performed using CellPhy v0.9.2[40], which utilizes RXML-NG[54], a maximum likelihood framework for phylogenetic inference. CellPhy estimates the allelic dropout rate and false positive rate per sample and constructs the most likely tree based on the phenotype likelihoods (PL) and these estimates. Before running CellPhy, the bulk t-MN samples were removed from the VCF. The bulk t-MN data was only used to validate the tree (see below). CellPhy was run on the PL with standard settings, including the "GT16+FO+E" model from CellPhy. Next, 100 bootstrap iterations were run. CellPhy was also used to map mutations to the tree (using the "mutmap" function) with the "--opt-branches off" setting.

Mutations of end branches were filtered out if one or more reads supporting the alternative allele was present in more than one cell. These could be mutations that were true in one cell, and were technical artifacts in other cells, but to make the final set more robust, these were filtered out. In addition, the end branches containing a single cell for which PTA WGA was used were filtered on the mutations that were part of the PTATO output. This was done to ensure no PTA-artifacts were included in the tree. Finally, the trees were rooted with treeio's "root" function, using the MSC cells as the outgroup, or an HSPC which shared no mutations with the other samples, when MSCs were unavailable (UPN008).

Fitting mutational signatures to the mutations of single branches of the tree was done with MutationalPattern's function "fit_to_signatures_strict" with option "max_delta=0.05". Trees were visualized using the R package "ggtree" v3.4.1[55] which uses "ape" v.5.6-2[56].

Analysis of a WGS cohort of metastases

A cohort of solid tumor metastases previously described by Priestley et al.[35] was used to verify signatures that were not described in the COSMIC database. Somatic and germline mutation calls as well as copy number status were obtained from the Hartwig Medical Foundation (HMF) that created this dataset. SBSG was refitted to the dataset together with COSMIC signatures v3.2 using the function "fit_to_signatures_strict" from the MutationalPatterns package with option "max_delta=0.01". Previously, Pich et al.[30] have extracted signatures from this cohort, including signature "sbs25", that was similar to the signature SBSD that we extracted from our cohort. We therefore refitted these signatures, including a signature similar to COSMIC SBS31, together with COSMIC SBS35 to this dataset in the same manner. In this case, only patients that were treated with cisplatin or carboplatin were selected based on the meta data of the data set.

TP53 mutation status was extracted from three different sources. First, a TP53 allele was considered to be lost if the BAF of the region including (a part of) TP53 was higher than 0.95, as determined by the above mentioned gridss-purple-linx pipeline from HMF. Somatic and germline calls were also obtained from HMF

and exonic TP53 mutations that were annotated in ClinVar as "Pathogenic" or "Likely Pathogenic" were considered driver mutations, as well as all other nonsense and frameshift mutations. For the somatic mutations, also all missense, structural interaction variants, splice site variants, and protein-protein interaction variants were considered driver mutations.

Cell culture and genotyping

The MV4-11^{bulk} and MV4-11^{R248W} cells were kindly provided by the Frank van Leeuwen and Willem Cox (Princess Máxima Center for pediatric oncology, Utrecht, The Netherlands). The frequency of the *TP53* R248W variant was determined by Sanger Sequencing using the Indigo tool (Gear-Genomics)[57]. MV4-11 cells were cultured at 37°C, 5%CO2 and replated biweekly at 4*10^5 cells/mL in IMDM (Gibco™), supplemented with 10% FCS (Sigma-Aldrich) and 1% Penicillin-Streptomycin (Gibco™). The subclonal and clonal, homozygous presence of the *TP53* R248W variant in MV4-11bulk and MV4-11R^{248W} respectively were confirmed by Sanger sequencing (Macrogen) following cell lysis (DirectPCR, Viagen Biotech) and amplification (GoTaq™ Hot Start Master Mix, Promega) using the following primers: Fw – CCTGCTTGCCACAGGTCTC; Rev – GGGGATGTGATGAGAGGTGG (IDT). Visualization of genotyping was performed through alignment to ENST00000269305.9 in Benchling (2023).

Carboplatin resistance assay

Following staining by CellTrace™ Far Red (Invitrogen™), MV4-11 cells were seeded at 25.000 cells per mL for a carboplatin (Fresenius Kabi) treatment range and an untreated condition. After four days, cell viability and proliferation were assessed by DAPI staining (Sigma-Aldrich) and subsequent flow cytometry (CytoFLEX S, Beckman Coulter). For cell viability, we determined the DAPI-negative fraction of single cells in FlowJo™ (v10.8.1) and extracted IC50 (ED50) values from the summary of the dose-response model using the R package DRC (v.3.0-1)[58]. Representative gating images are available in **Fig. S6**. The proliferation score was calculated as follows. Per condition, the median fluorescent CellTrace™ intensity (MFI) of DAPI-negative cells was determined. Next, per condition, the MFI was normalized to the MFI at the start of treatment, and then to the untreated condition at day four. Subsequently, the normalized CellTrace™ intensities were rescaled to a range of 0-1 and transformed (*ProliferationScore*=1-MFI$_{scaled}$). Treatment conditions above 10μM carboplatin were not taken along as the number of viable cells was insufficient (<1.000 cells) for CellTrace™ measurements. Each experiment consisted of the average of three technical replicates per condition and was performed independently three times. Visualization was performed using the R packages DRC (v.3.0-1)[58] and ggpubr (v.0.6.0)[59].
The fitted dose-response curves were tested by ANOVA against the null model, which lacked the genotype factor of MV4-11 bulk and R248W (R package stats v4.2.2). The IC50 values were compared by z-test (default settings DRC::compParm v.3.0-1)[58]. The Proliferation Scores were compared using a one-sided t-test and holm's multiple testing correction (rstatix v.0.7.2)[60].

Data visualization

All data visualization that was not performed with the packages mentioned above was performed using the R package "ggplot2" v.3.3.6, which is part of the "tidyverse" suite of packages[61].

Data availability

The whole genome sequencing data generated for this study are available at the European Genome-phenome Archive (EGA, www.ebi.ac.uk/ega/) under accession number EGAS00001005141. The filtered VCF files are available at Mendeley Data (https://data.mendeley.com/datasets/72cvzs5dg7/draft?a=739eb6f0-c243-427a-8b59-7fb394614d35).

Acknowledgements

This work was funded by an ERC Consolidator grant to R. van Boxtel from the European Research Council (ERC; no. 864499). In addition, this work was supported by the Oncode institute, funding E.J.M. Bertrums, J.K. de Kanter, M. Verheul, M.J. van Roosmalen, and R. van Boxtel. The authors want to thank the Hartwig Medical Foundation for facilitating the low-input whole-genome sequencing. The authors thank Willem Cox, Frank van Leeuwen and Ronald Stam for kindly providing the MV4-11 cell lines. Ruben van Boxtel is a New York Stem Cell Foundation – Robertson Investigator. This research was supported by The New York Stem Cell Foundation. The authors want to acknowledge members of the International Berlin-Frankfurt-Münster AML Study Group (I-BFM AML SG) for the inclusion of patients in this study.

Author contributions

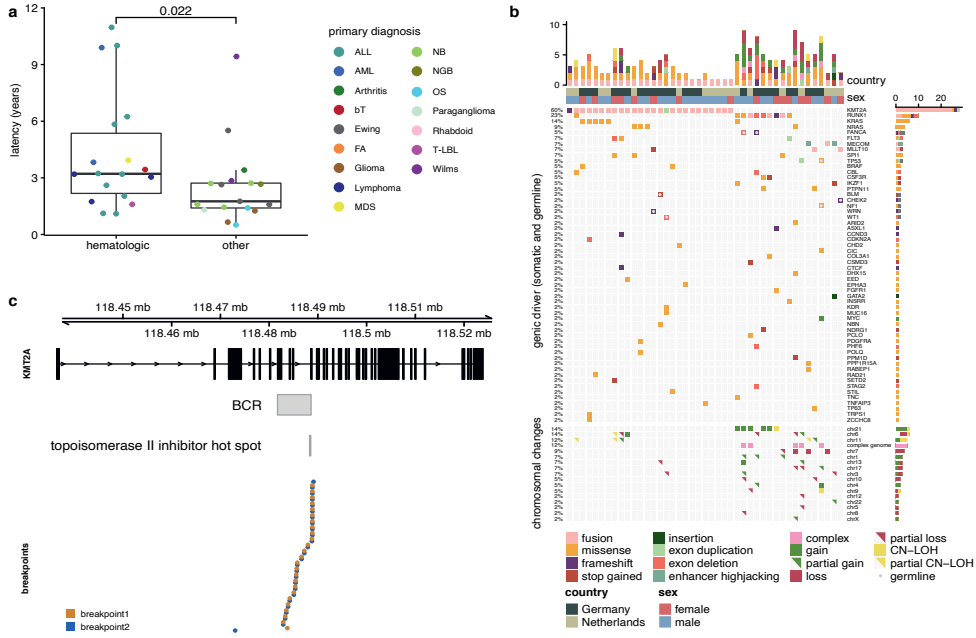
E.J.M.B. and J.K.d.K were responsible for the experimental design. E.A. and D.R. included patient samples and assembled clinical data. H.H., D.R., B.F.G., C.M.Z. and R.v.B. supported the international collaborative study protocol. E.J.M.B. performed all experimental work, with support of M.V. L.L.M.D. performed the *in vitro* validation experimental work, with support of L.T. J.K.d.K performed all bioinformatic data analysis, with support of M.J.vR. E.J.M.B. and J.K.d.K were responsible for data interpretation. E.J.M.B., J.K.d.K, L.M.M.D. and R.v.B drafted the manuscript. All authors have proof-read the manuscript.

References

1. Hurley, L. H. DNA and its associated processes as targets for cancer therapy. Nat Rev Cancer 2, 188–200 (2002).
2. Voso, M. T., Falconi, G. & Fabiani, E. What's new in the pathogenesis and treatment of therapy-related myeloid neoplasms. Blood 138, 749–757 (2021).
3. McNerney, M. E., Godley, L. A. & Le Beau, M. M. Therapy-related myeloid neoplasms: When genetics and environment collide. Nat Rev Cancer 17, 513–527 (2017).
4. Teepen, J. C. et al. Long-Term Risk of Subsequent Malignant Neoplasms After Treatment of Childhood Cancer in the DCOG LATER Study Cohort: Role of Chemotherapy. Journal of Clinical Oncology 35, 2288–2298 (2017).
5. Aguilera, D. G. et al. Pediatric Therapy-related Myelodysplastic Syndrome/Acute Myeloid Leukemia. J Pediatr Hematol Oncol 31, 803–811 (2009).
6. Kayser, S. et al. The impact of therapy-related acute myeloid leukemia (AML) on outcome in 2853 adult patients with newly diagnosed AML. Blood 117, 2137–2145 (2011).
7. Wong, T. N. et al. Role of TP53 mutations in the origin and evolution of therapy-related acute myeloid leukaemia. Nature 518, 552–555 (2015).
8. Coombs, C. C. et al. Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. Cell Stem Cell 21, 374-382.e4 (2017).
9. Pich, O. et al. The evolution of hematopoietic cells under cancer therapy. Nat Commun 12, 4803 (2021).
10. Bolton, K. L. et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nat Genet 52, 1219–1226 (2020).
11. Wong, T. N. et al. Cellular stressors contribute to the expansion of hematopoietic clones of varying leukemic potential. Nat Commun 9, 455 (2018).
12. Diamond, B. et al. Tracking the evolution of therapy-related myeloid neoplasms using chemotherapy signatures. Blood 141, 2359–2371 (2023).
13. Hagiwara, K. et al. Dynamics of Age- versus Therapy-Related Clonal Hematopoiesis in Long-term Survivors of Pediatric Cancer. Cancer Discov 13, 844–857 (2023).
14. Bertrums, E. J. M. et al. Elevated Mutational Age in Blood of Children Treated for Cancer Contributes to TherapyRelated Myeloid Neoplasms. Cancer Discov 12, 1860–1872 (2022).
15. Schwartz, J. R. et al. The acquisition of molecular drivers in pediatric therapy-related myeloid neoplasms. Nat Commun 12, 985 (2021).
16. Spitzer, B. et al. Bone marrow surveillance of pediatric cancer survivors identifies clones that predict therapy-related leukemia. Clinical Cancer Research clincanres.2451.2021 (2022) doi:10.1158/1078-0432.CCR-21-2451.
17. Coorens, T. H. H. et al. Clonal hematopoiesis and therapy-related myeloid neoplasms following neuroblastoma treatment. Blood 137, 2992–2997 (2021).
18. Le, H. et al. Rearrangements of the MLL gene are influenced by DNA secondary structure, potentially mediated by topoisomerase II binding. Genes Chromosomes Cancer 48, 806–815 (2009).
19. Mirault, M.-E., Boucher, P. & Tremblay, A. Nucleotide-Resolution Mapping of Topoisomerase-Mediated and Apoptotic DNA Strand Scissions at or near an MLL Translocation Hotspot. The American Journal of Human Genetics 79, 779–791 (2006).
20. Preudhomme, C. et al. High incidence of biallelic point mutations in the Runt domain of the AML1/PEBP2αB gene in Mo acute myeloid leukemia and in myeloid malignancies with acquired trisomy 21. Blood 96, 2862–2869 (2000).
21. Bolouri, H. et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat Med 24, 103–112 (2018).
22. Gröbner, S. N. et al. The landscape of genomic alterations across childhood cancers. Nature 555, 321–327 (2018).
23. Zhang, J. et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. New England Journal of Medicine 373, 2336–2346 (2015).
24. Baranwal, A., Hahn, C. N., Shah, M. V. & Hiwase, D. K. Role of Germline Predisposition to Therapy-Related Myeloid Neoplasms. Curr Hematol Malig Rep 17, 254–265 (2022).
25. Osorio, F. G. et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep 25, 2308-2316.e4 (2018).
26. Machado, H. E. et al. Diverse mutational landscapes in human lymphocytes. Nature 608, 724–732 (2022).
27. Boot, A. et al. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. Genome Res 28, 654–665 (2018).
28. Li, B. et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. Blood 135, 41–55 (2020).

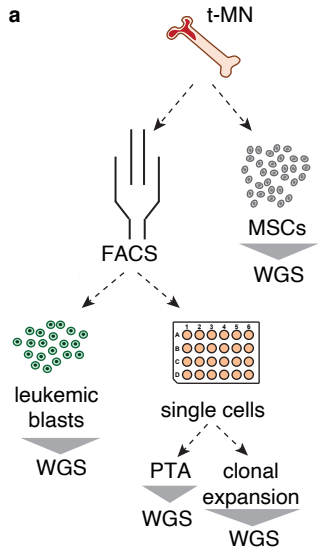
29. Christensen, S. et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun 10, 4571 (2019).
30. Pich, O. et al. The mutational footprints of cancer therapies. Nat Genet 51, 1732–1740 (2019).
31. Smith, H. L., Southgate, H., Tweddle, D. A. & Curtin, N. J. DNA damage checkpoint kinases in cancer. Expert Rev Mol Med 22, (2020).
32. Baugh, E. H., Ke, H., Levine, A. J., Bonneau, R. A. & Chan, C. S. Why are there hotspot mutations in the TP53 gene in human cancers? Cell Death Differ 25, 154–160 (2018).
33. Zhou, X., Hao, Q. & Lu, H. Mutant p53 in cancer therapy-the barrier or the path. J Mol Cell Biol 11, 293–305 (2019).
34. Bordin, F. et al. WT1 loss attenuates the TP53-induced DNA damage response in T-cell acute lymphoblastic leukemia. Haematologica 103, 266–277 (2018).
35. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature 575, 210–216 (2019).
36. Kucab, J. E. et al. A Compendium of Mutational Signatures of Environmental Agents. Cell 177, 821-836.e16 (2019).
37. Brady, S. W. et al. The Clonal Evolution of Metastatic Osteosarcoma as Shaped by Cisplatin Treatment. Mol Cancer Res 17, 895–906 (2019).
38. Morganella, S. et al. The topography of mutational processes in breast cancer genomes. Nat Commun 7, 11383 (2016).
39. Gonzalez-Pena, V. et al. Accurate genomic variant detection in single cells with primary template-directed amplification. Proc Natl Acad Sci U S A 118, e2024176118 (2021).
40. Kozlov, A., Alves, J. M., Stamatakis, A. & Posada, D. CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. Genome Biol 23, 37 (2022).
41. Bowman, R. L., Busque, L. & Levine, R. L. Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. Cell Stem Cell 22, 157–170 (2018).
42. Light, N. et al. Germline TP53 mutations undergo copy number gain years prior to tumor diagnosis. Nat Commun 14, 77 (2023).
43. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. Nature 578, 94–101 (2020).
44. Petljak, M. & Maciejowski, J. Molecular origins of APOBEC-associated mutations in cancer. DNA Repair (Amst) 94, 102905 (2020).
45. Yan, B. et al. Low-frequency TP53 hotspot mutation contributes to chemoresistance through clonal expansion in acute myeloid leukemia. Leukemia 34, 1816–1827 (2020).
46. Willis, A., Jung, E. J., Wakefield, T. & Chen, X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. Oncogene 23, 2330–2338 (2004).
47. Mitchell, E. et al. Clonal dynamics of haematopoiesis across the human lifespan. Nature 606, 343–350 (2022).
48. Huang, Z. et al. Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. Nat Genet 54, 492–498 (2022).
49. Middelkamp, S. et al. Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox. Cell genomics 3, 100389 (2023).
50. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849 (2016).
51. Danecek, P. et al. Twelve years of SAMtools and BCFtools. Gigascience 10, giab008 (2021).
52. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
53. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. Bioinformatics 30, 2811–2812 (2014).
54. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35, 4453–4455 (2019).
55. Xu, S. et al. Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. iMeta 1, e56 (2022).
56. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526–528 (2018).
57. Rausch, T., Fritz, M. H.-Y., Untergasser, A. & Benes, V. Tracy: basecalling, alignment, assembly and deconvolution of sanger chromatogram trace files. BMC Genomics 21, 230 (2020).
58. Ritz, C., Baty, F., Streibig, J. C. & Gerhard, D. Dose-Response Analysis Using R. PLoS One 10, e0146021–e0146021 (2015).
59. Alboukadel Kassambara. ggpubr: 'ggplot2' Based Publication Ready Plots. https://rpkgs.datanovia.com/ggpubr/ (2023).
60. Alboukadel Kassambara. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. https://rpkgs.datanovia.com/rstatix/ (2023).
61. Wickham, H. et al. Welcome to the Tidyverse. J Open Source Softw 4, 1686 (2019).

Supplementary Material

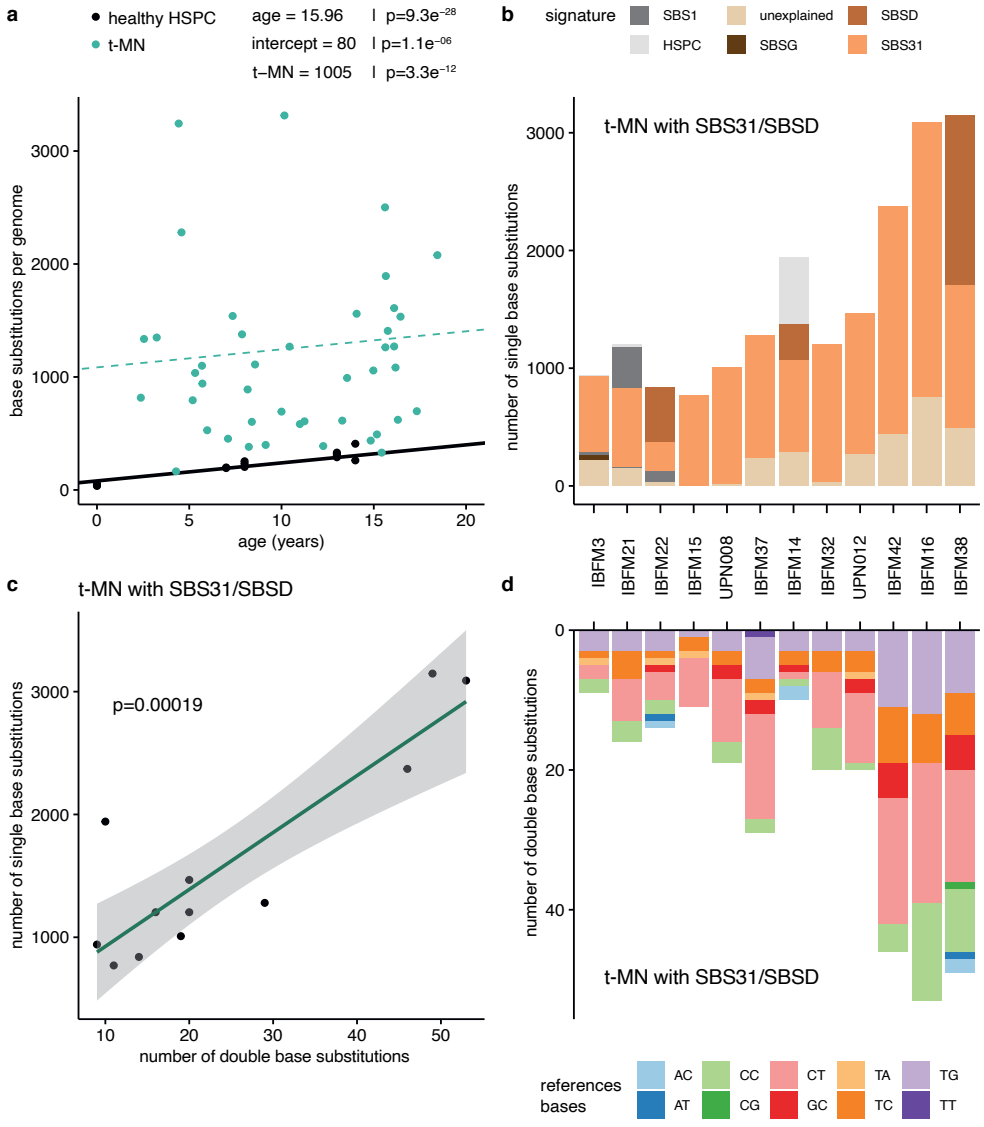


Supplementary Figure 1. Additional cohort details

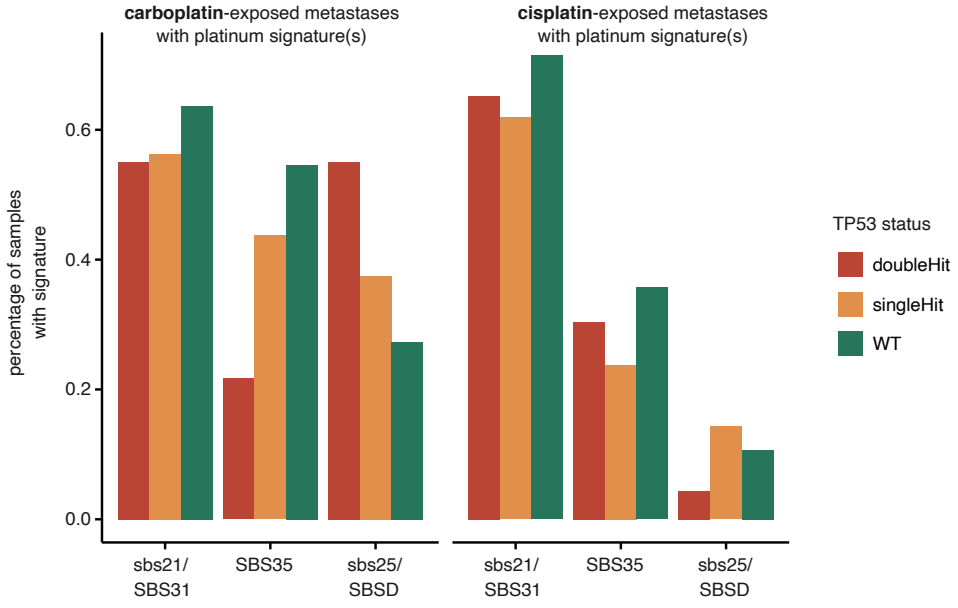
a) Boxplot depicting the latency time in years between the first diagnosis and t-MN development. Colors represent first diagnosis. ALL: acute lymphoblastic leukemia; AML: acute myeloid leukemia; bT: beta-thalassemia; FA: Fanconi anemia; MDS: myelodysplastic syndrome; NB: neuroblastoma; NGB: neuroganglioblastoma; OS: osteosarcoma; SCT: allogenic stem cell transplantation; TLBL: T-cell lymphoblastic lymphoma. Two-sided Wilcox-test. **b)** Oncoprint with all identified driver mutations in all t-MN. In contrast to main Fig. 1e also mutations found in single t-MN were included. The bar plots on top represent the number of driving events present in each sample. The bar plots on the right represent the number of patients with the driver. CN-LOH: copy neutral loss of heterozygosity. **c)** *KMT2A* (MLL) breakpoints from t-MN patients in our cohort. Indicated on top are the general *KMT2A* breakpoint cluster region (BCR) and the hot spot associated with topoisomerase II inhibitors (TOP2i).



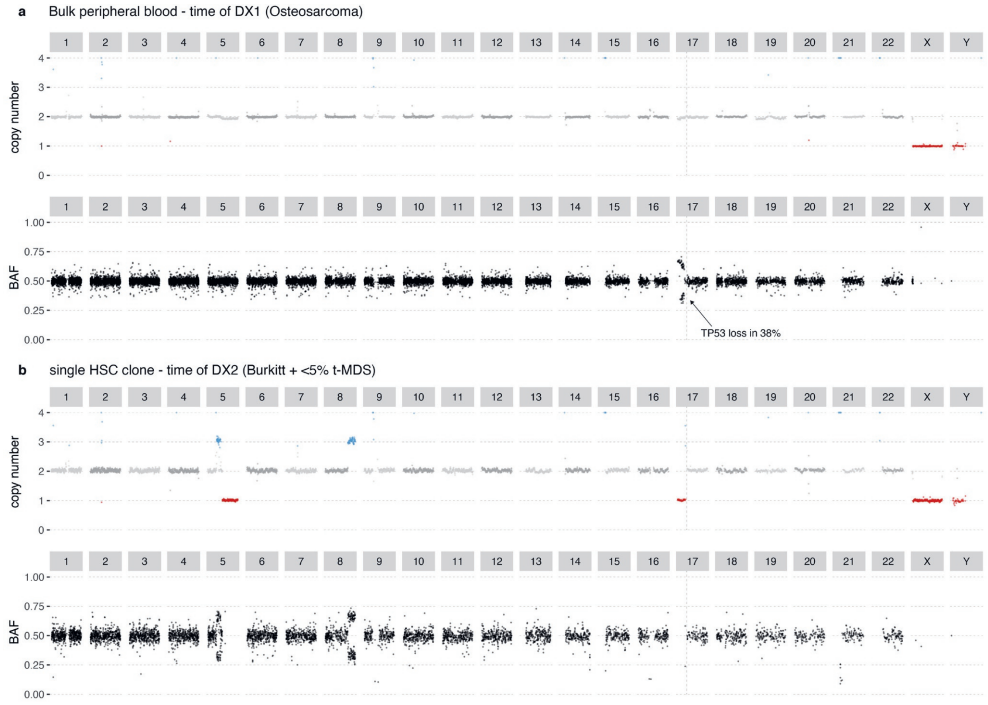
Supplementary Figure 2. Experimental setup
A schematic overview of the experimental setup of this study. In short, bone marrow biopsies at time of t-MN were collected. Blasts and HSPCs were purified by FACS. Blasts were sorted in bulk and single cell in a 96-wells plate for primary template-directed amplification (PTA). Single HSPCs were sorted in a 384-wells plate for clonal expansion and in a 96-wells plate for PTA. Mesenchymal stromal cells (MSCs) were plated and expanded *in vitro*.



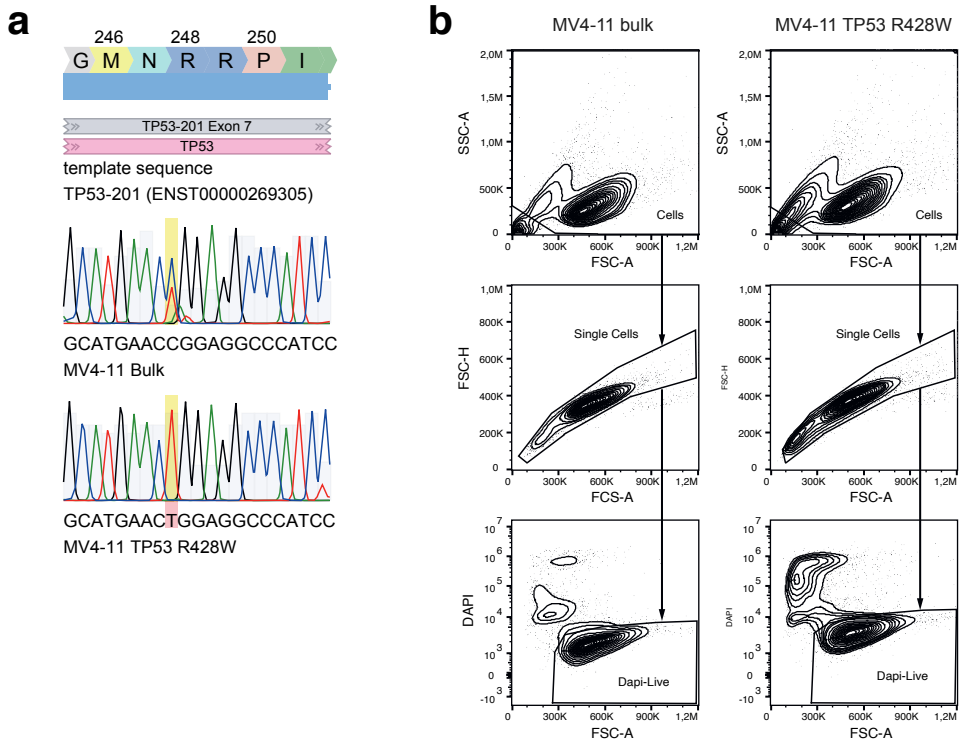
Supplementary Figure 3. Mutation burden of patient with platinum-related mutations
a) Mutation accumulation of t-MN (colored dots) compared to the baseline of healthy hematopoietic stem and progenitor cells (HSPCs; black dots). A linear mixed-effects model was run on both the baseline and t-MN data, taking into account the donor, the age and the mutation load. The effects for age (per year), the intercept of the baseline with the y-axis (number of mutations at birth), and the additional effect for t-MN are stated, together with p-values. t-MN on average had 1005 additional mutations compared to the healthy baseline with a p-value of $3.3*10^{-12}$. Conditional R^2=0.998. **b)** The signature contribution of t-MN that harbored a contribution of the platinum-related signatures SBS31 and/or SBSD. **c)** A linear model of the single base and double base substitutions in the t-MN patients depicted in (b) where n_snv = 461 + 46 * n_dbs, adjusted R^2=0.74. **d)** The number and type of double base substitutions of the same t-MN depicted in (b). These are similar to DBS5, the COSMIC double base substitutions signature linked to platinum-based compound exposure.



Supplementary Figure 4: Platinum-related signatures in metastases by TP53 status. Carboplatin and cisplatin exposed metastases from a previously described cohort by Priestley et al.[36] that displayed any cisplatin-related signatures (SBSD/sbs25, SBS31/sbs21, SBS35) after bootstrapped refitting (n=100) were included. The bars represent the percentage of samples that had contribution of each signature. The samples were split by TP53 status: wild-type (WT), single hit, double hit (most often a somatic mutation and a loss of the wild-type allele).



Supplementary Figure 5. WGS data of patient UPN034 at first diagnosis and t-MN.
a) Copy number plot and B allele frequency plot of whole genome sequencing (WGS) data of peripheral blood of patient UPN034 at the time of osteosarcoma (DX1). Data from the diagnostics department of our institute. **b)** Similar to (a) but for a bone marrow HSPC at time of DX2.



Supplementary Figure 6.
a) Genotyping results of MV4-11bulk (middle) and MV4-11R248W (bottom), aligned to ENST00000269305.9 (top). The nucleotide variant is highlighted in yellow. **b)** Representative gating of cells (top), single cells (middle), and DAPI-negative cells (bottom) for MV4-11bulk (left) and MV4-11R248W (right).

Supplementary Table 1. Clinical patient and treatment information.

Patient ID	F/M	Primary cancer / treatment protocol	Age Dx1 (y)	Chemotherapy agents (from databases or extracted from protocol)	SCT	RT	Age t-MN (y)
IBFM01	M	ALL | ALL-BFM-2000	2.3	Vincristine, Daunorubicin, Asparaginase, MTX, Cyclophosphamide, 6-MP, ARA-C, Doxorubicin, 6-TG. If HR: + Ifosfamide, Vindesine	NA	NA	8.6
IBFM02	M	ALL | ALL-BFM-2000	3.7	Vincristine, Daunorubicin, Asparaginase, MTX, Cyclophosphamide, 6-MP, ARA-C, Doxorubicin, 6-TG. If HR: + Ifosfamide, Vindesine	No	No	5.7
IBFM03	F	Osteosarcoma | NA	NA	NA	NA	NA	13.6
IBFM07	F	ALL | ALL-BFM-92	5.2		Yes	NA	16.2
IBFM09	F	ALL | NA	3.5	NA	NA	NA	4.6
IBFM10	M	Ewing Sarcoma | Euro-Ewing 99	14.5	Vincristine, Ifosfamide, Doxorubicin, Etoposide, Actinomycin D. If R1: potentially + Cyclophosphamide. If R2: potentially + Busulfan, Melphalan. If R3: potentially + Melphalan or Treosulfan/Melphalan or Busulphan/Melphalan	NA	NA	16.1
IBFM11	M	MDS | NA	12.4	NA	No	NA	16.3
IBFM14	M	Ewing Sarcoma | Euro-Ewing 99	12.9	Vincristine, Ifosfamide, Doxorubicin, Etoposide, Actinomycin D. If R1: potentially + Cyclophosphamide. If R2: potentially + Busulfan, Melphalan. If R3: potentially + Melphalan or Treosulfan/Melphalan or Busulphan/Melphalan	NA	NA	18.4
IBFM15	F	Neuroblastoma (St. IV) Studytherapy NB 2004 (HR) + 2x N8	3.7	Cisplatin, Etoposide, Vindesin, Vincristine, Dacrabaxine, Ifosfamide, Doxorubicin	Yes	NA	5.2
IBFM16	F	Neuroblastoma (St. III) NB2004	2.8	Cisplatin, Etoposide, Vindesin, Vincristine, Dacarbazine, Ifosfamide, Doxorubicin, 13-cis-retinoic acid. If MR: + Cyclophosphamide	No[1]	Yes1	4.4
IBFM21	M	Paraganglioma GPOH MET 97		Vincristine, Ifosfamide, Doxorubicin, Carboplatin, Etoposide	No[1]	No	10.4
IBFM22	F	Opticus glioma SIOP-LGG-2004 (NF1)	7.5	Vincristine, Carboplatin (if allergic Cisplatin and Cyclophosphamide)	No[1]	NA	8.2
IBFM25	M	Nephroblastoma SIOP-2001	5.4	Actinomycin D, Vincristine. Potentially: + Doxorubicin, Etoposide, Carboplatin, Cyclophosphamide,	No[1]	NA	8.2
IBFM26	M	Fanconi anemia | NA	NA	NA	NA	NA	14.8
IBFM27	M	Hodgkin lymphoma| NA	5.3	NA	NA	NA	8.4
IBFM28	M	Atypical rhabdoid tumor | NA	NA	NA	NA	NA	2.5
IBFM29	F	Idiopathic arthritis | Immunosuppressives	0.9	NA	NA	NA	4.3
IBFM31	M	c-ALL | AIEOP BFM-ALL 2009 HR	13.3	Vincristine, Daunorubicin, MTX, Aparaginase, Cyclophosphamide, ARA-C, 6-MP, Vindesine, Ifosfamide, Etoposide, Doxorubicin, 6-TG. Potentially: + Fludarabine, Daunoxome	No	No	16.4
IBFM32	M	(Osteosarcoma &) Pre-B-ALL | AIEOP BFM-ALL 2009 HR	15.1	(Osteosarcoma treatment: NA). Vincristine, Daunorubicin, MTX, Aparaginase, Cyclophosphamide, ARA-C, 6-MP, Vindesine, Ifosfamide, Etoposide, Doxorubicin, 6-TG. Potentially: + Fludarabin, Daunoxome	No	No	15.6
IBFM33	M	Intrathoracic Ewing sarcoma |Ewing 2008	13.5	Vincristine, Ifosfamide, Doxorubicin, Etoposide, Actinomycin D. If R1: potentially + Cyclophosphamide. If R2: potentially + Busulfan, Melphalan. If R3: + Cyclophosphamide, potentially + Treosulfan/Melphalan	No	Yes	16.1
IBFM35	M	Fanconi Anemia |NA	NA	NA	Yes	NA	17.3

Supplementary Table 1. *continued*

Patient ID	F/M	Primary cancer / treatment protocol	Age Dx1 (y)	Chemotherapy agents (from databases or extracted from protocol)	SCT	RT	Age t-MN (y)
IBFM36	F	Pre-B-ALL |AIEOP BFM-ALL 2009	NA	Vincristine, Daunorubicin, MTX, Asparaginase, Cyclophosphamide, ARA-C, 6-MP, Leukovorin, Doxorubicin, 6-TG If HR: + Vindesine, Ifosfamide, Etoposide, Doxorubicin, potentially + Fludarabine/ Daunoxome	No	NA	2.4
IBFM37	F	Neuroblastoma | NA	0.5	NA	NA	NA	3.2
IBFM38	M	Nephrobastoma | SIOP-2001	0.7	Actinomycin D, Vincristine. Potentially: + Doxorubicin, Etoposide, Carboplatin, Cyclophosphamide	No	No	10.2
IBFM42	M	Osteosarcoma | NA	NA	NA	NA	NA	15.6
IBFM43	M	ALL | AIEOP BFM-ALL 2009	4.2	Vincristine, Daunorubicin, MTX, Asparaginase, Cyclophosphamide, ARA-C, 6-MP, Leukovorin, Doxorubicin, 6-TG. If HR: + Vindesine, Ifosfamide, Etoposide, Doxorubicin, potentially.+ Fludarabine/ Daunoxome	No	No	5.9
UPN001	F	Burkitt |LMB 2001	11.5	Cyclophosphamide, ARA-C, MTX, Doxorubicin, Vincristine, Etoposide	No	No	13.2
UPN002	F	B-ALL	3.9	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase	No	No	5.0
UPN003	M	ALL | ALL10, ALL-R3, ALL 11 HR + AD-HOC	5.7	At time of FU: MTX, Vincristine, ARA-C, PEG-asparaginase, Daunorubicin, Cyclophosphamide, 6-MP, Doxorubicin, Mitoxantrone, 6-TG. After FU new: ATG, BuFluClo, Teniposide, allogenic MUD-SCT	Yes, 2x	NA	15.7
UPN004	M	B-ALL	15.5	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase, Cyclophosphamide, 6-MP, Doxorubicin	No	No	16.7
UPN005	M	B-ALL	4.3	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase, Cyclophosphamide, 6-MP, Doxorubicin	No	No	5.1
UPN006	F	B-ALL	3.4	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase, Cyclophosphamide, 6-MP	No	No	3.6
UPN007	F	B-ALL	8.3	Vincristine, Daunorubicin, MTX, ARA-C, Asparaginase, Cyclophosphamide, 6-MP, Doxorubicin	No	No	9.4
UPN008	F	Osteosarcoma | Euramos1	13.5	Doxorubicin, Cisplatin, MTX. If randomized/poor response: + Ifosfamide, Etoposide	NA	NA	14.9
UPN009	M	Non low-grade astriocytoma |SIOP LGG2004	14.1	None	No	Yes	15.4
UPN010	M	Ewing | Ewing2008R3	3.9	Ifosfamide, Doxorubicin, Actinomycin D, Vincristine, Cyclophosphamide, Etoposide	No	Yes	5.7
UPN011	F	Neuro-ganglioblastoma | NBL2009MRG	3.3	Cisplatin, Etoposide, Vindesine, Vincristine, Dacarbazine, Ifosfamide, Doxorubicin, low dose Cyclophosphamide, Retinoic acid	No	Yes	5.9
UPN012	M	Neuroblastoma	4.6	Cisplatin or carboplatin, etoposide, Vindesine, Dacarbazine, Doxorubicin, Ifosfamide, Vincristine, Busulfan, Melfalan.	No	Yes	7.4
UPN013	M	b-thalassemia | NA	NA	Treosulfan, Fludarabine, Thiotepa, ATG, Alemtuzumab	Yes, 2x	NA	5.3
UPN014	M	ALL | ALL11-MRG	6.0	Vincristine, Daunorubicin, Asparaginase, Cyclophosphamide, ARA-C, 6-MP, MTX, Doxorubicin	No	No	7.1
UPN015	F	Lymphoma | NA	NA	NA	NA	NA	15.6
UPN016	F	NHL | ALL VII	NA	Daunorubicin, 6-TG, Vindesine, 6-MP, Asparaginase, Cyclophosphamide, Vincristine, Doxorubicin, Teniposide, MTX, ARA-C, Ifosfamide	No	No	11.2
UPN017	M	AML | ANLL92	5.2	Doxorubicin, Cyclophosphamide, Idarubicin, ARA-C, Vincristine, Mitoxantrone, Etoposide, 6-TG	No	No	15.1
UPN018	M	pre-B ALL | ALL8-MRG	4.6	Vincristine, Daunorubicin, Asparaginase, ARA-C, MTX, 6-MP, Doxorubicin, Cyclophosphamide, 6-TG	No	No	7.8

Supplementary Table 1. *continued*

Patient ID	F/M	Primary cancer / treatment protocol	Age Dx1 (y)	Chemotherapy agents (from databases or extracted from protocol)	SCT	RT	Age t-MN (y)
UPN019	M	AML | ANLL94	7.1	ARA-C, Idarubicin, Etoposide, Mitoxantrone, 6-TG, Vincristine, Doxorubicin, Cyclophosphamide. Conditioning: + Busulfan	Yes	TBI	10.9
UPN020	M	ALL | NA	8.2	NA	NA	NA	14.0
UPN022	F	Ewing sarcoma | NA	NA	NA	NA	NA	9.1
UPN023	F	T-ALL | ALL10-MRG	9.6	Vincristine, Daunorubicin, Asparaginase, Cyclophosphamide, 6-MP, ARA-C, Leukovorin, MTX, Doxorubicin	No	No	12.2
UPN024	M	T-LBL | Euro-LB02 III/IV	8.4	Vincristine, Daunorubicin, MTX, Asparaginase, 6-MP, ARA-C, Cyclophosphamide, Doxorubicin, 6-TG	No	No	10.0
UPN034	M	Osteosarcoma | Euramos1	15.7	Doxorubicin, Cisplatin, MTX	No	No	15.9

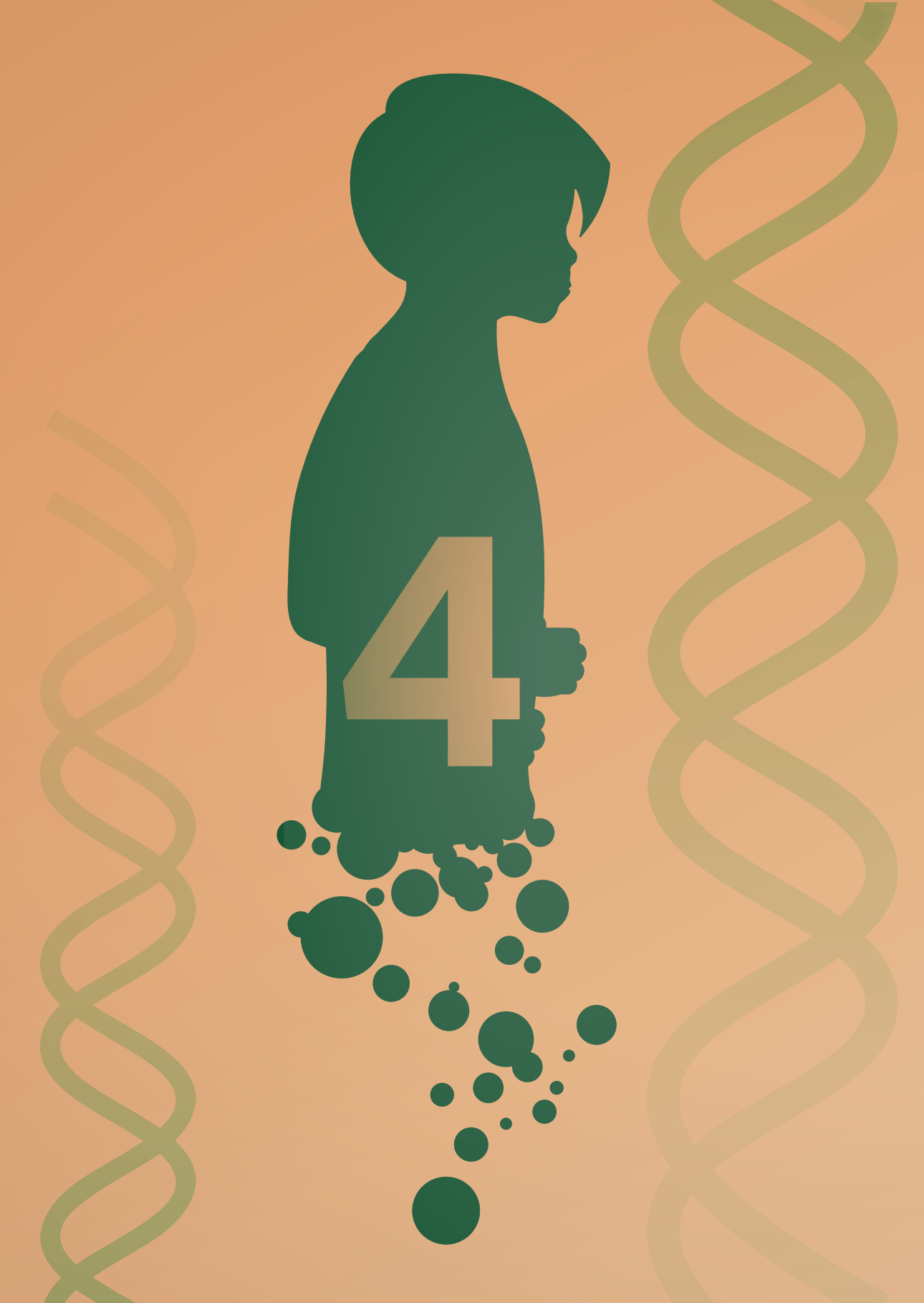
All chemotherapy has been retrieved from clinical data or extracted to our best knowledge from stated protocols, this can vary per individual patient. [1]According to treatment protocol. 6-MP = Mercaptopurine; 6-TG = thioguanine; ALL = acute lymphoblastic leukaemia; alloSCT = allogenic stem cell transplantation; AML = acute myeloid leukaemia; ARA-C = cytarabin; BuFluClo = Busulfan, Fludarabin, Clofarabin; Dx1 = first diagnosis; F = female; HR = high risk; M = male; MRG = medium risk group; MTX = methotrexate; MUD = matched-unrelated-donor; NA = not available; NF1 = neurofibromatosis type 1; NHL = non-Hodgkin lymphoma; RT = radiotherapy; SCT = stem cell transplantation; St = stage; TBI = total body irradiation; t-MN = therapy-related myeloid neoplasm; y = years.

Supplementary Table 2. Cancer drivers and mutational signature contributions per patient.

patient ID	small drivers	fusions	SBS	indels	DBS	SBSE	SBSA	HSPC	SBSB	SBSD	SBSF	SBS31	SBS87	SBS1	SBSC	SBSG	SBS17a	SBS17b	SBS5
IBFM10	STIL | BRAF	MLLT3-KMT2A	1596	55	3	0	0	638	0	0	0	0	0	0	0	632	0	0	38
IBFM14	NRAS | DHX15 ARID2 | PPM1D		2030	68	10	0	0	565	0	310	0	781	0	0	0	0	0	0	5
IBFM22	CIC		921	53	14	0	0	0	0	468	0	250	0	94	0	0	0	0	0
IBFM42	EED	THAP12-KMT2A	2636	173	46	0	0	0	0	0	0	1929	0	0	0	0	0	0	0
IBFM35	RUNX1		713	47	4	0	0	22	0	0	0	0	0	198	0	0	0	0	388
IBFM11	FGFR1 | ASXL1 RUNX1		649	53	2	0	0	241	0	0	0	0	0	163	0	0	0	0	121
IBFM36	NRAS	AFF1-KMT2A	861	79	3	0	0	0	3	0	0	0	4	182	0	9	0	0	473
IBFM37		AFDN-KMT2A	1376	38	29	0	0	0	0	0	0	1045	0	0	0	0	0	0	0
IBFM1	CSF3R | NDRG1 PTPN11 | RUNX1		1109	46	3	0	0	171	0	0	0	0	454	259	0	9	0	0	28
IBFM2	RAD21 KRAS	ERBB4-KMT2A	1078	32	1	0	0	0	0	0	0	0	823	157	0	0	0	0	3
IBFM3	MLLT10	MLLT3-KMT2A	1018	59	9	1	0	8	0	0	0	650	0	27	0	36	0	0	0
IBFM7	CSF3R COL3A1	MIR99 AHG-RUNX1	1076	39	4	0	0	183	0	0	0	0	195	390	0	0	0	0	10
IBFM16	NRAS | POLQ PDGFRA	MLLT3-KMT2A	3273	77	53	0	0	0	0	0	0	2335	0	0	0	0	0	0	0
IBFM26	RUNX1 | CBL STAG2 | PHF6		480	62	1	0	0	22	0	0	0	0	16	139	0	14	0	0	186
IBFM9	TP63 PTPN11	MLLT10-PICALM	2219	48	5	0	0	0	0	0	0	0	645	0	0	0	59	536	819
IBFM15	EPHA3	DCP1A-KMT2A	794	16	11	0	0	0	0	0	0	771	0	0	0	0	0	0	0
IBFM28	BRAF KRAS	MLLT3-KMT2A	1325	29	11	878	0	3	0	124	20	4	0	0	0	0	0	0	0
IBFM29	NBN	KMT2A-MLLT1	177	21	0	0	0	12	0	0	0	1	8	40	0	19	0	0	51
IBFM25		MLLT3-KMT2A	396	30	1	0	0	57	0	0	0	0	2	114	1	6	0	0	139
IBFM27	CCND3 FLT3 CTCF	KMT2A-ELL	612	31	3	40	0	61	0	2	6	0	0	117	0	7	0	0	301
IBFM31	INSRR RUNX1 FLT3		1537	52	11	0	0	0	0	0	0	0	1534	0	0	0	0	0	0
IBFM33	RABEP1 PP1R15A MECOM		1248	45	2	0	0	4	9	0	31	14	0	20	17	785	0	0	67
IBFM21	SETD2 | SPI1 FLT3	MLLT3-KMT2A	1308	72	16	0	0	25	0	0	350	674	0	0	0	0	0	0	6
IBFM32	TRPS1 | KRAS ZCCHC8 CDKN2A	AFF1-KMT2A	1292	48	20	0	0	0	0	0	0	1174	0	0	0	0	0	0	0
IBFM38	KDR | MUC16 KMT2A		3356	111	49	0	0	0	0	1444	0	1214	0	0	0	0	0	0	0
UPN001	MLLT10	AFDN-KMT2A	626	42	1	4	0	394	1	0	0	0	0	37	5	23	0	0	57
UPN003	TNFAIP3	MLLT3-KMT2A	1399	48	5	6	329	229	0	0	0	0	328	62	0	7	0	0	286
UPN008	CHD2	MLLT3-KMT2A	1084	37	19	0	0	0	0	0	0	992	0	0	0	0	0	0	0

Supplementary Table 2. *continued*

patient ID	small drivers	fusions	SBS	indels	DBS	SBSE	SBSA	HSPC	SBSB	SBSD	SBSF	SBS31	SBS87	SBS1	SBSC	SBSG	SBS17a	SBS17b	SBS5
UPN009		KMT2A-MLLT6	325	15	0	0	0	241	0	0	0	11	11	10	0	4	0	0	15
UPN010	KRAS | KRAS	KMT2A-MLLT1	946	41	4	1	0	299	0	0	5	7	0	31	0	361	0	0	75
UPN011		MLLT10-DDX3X	570	54	7	4	0	117	0	0	0	61	4	0	2	83	0	0	134
UPN024		MLLT3-KMT2A	690	28	1	1	0	450	0	0	0	0	162	4	0	8	0	0	0
UPN012	IKZF1 | TNC KRAS | RUNX1		1546	39	20	0	0	0	0	0	0	1199	0	0	0	0	0	0	0
UPN013	CBL | RUNX1 RUNX1	KMT2A-SEPTIN9	1049	68	1	0	0	14	85	0	0	3	0	97	673	0	0	0	60
UPN014	KRAS	MLLT3-KMT2A	456	22	1	0	0	70	0	0	0	0	177	73	0	0	0	0	118
UPN015		AFDN-KMT2A	1925	130	4	0	0	5	1391	0	0	0	0	0	0	0	0	0	113
UPN016	SPI1	CBFA2T3-RUNX1	601	28	1	0	0	64	0	0	0	0	219	58	0	40	0	0	137
UPN017		SATB1-MECOM	488	25	0	0	0	133	0	0	0	0	38	105	0	7	0	0	177
UPN018	KMT2A KMT2A	EPS15-KMT2A	1498	172	6	0	0	0	0	0	0	0	1357	0	0	0	0	0	0
UPN019	GATA2 | IKZF1 MECOM		636	79	1	0	0	10	0	0	0	0	4	165	0	0	0	0	331
UPN020	PCLO | CSMD3	RUNX1-HSF2BP	1696	204	11	0	0	0	0	0	0	0	361	432	0	0	0	0	549
UPN022		MLLT3-KMT2A	413	38	1	1	0	209	0	0	3	3	0	16	1	21	0	0	61
UPN023	TP53		419	44	2	0	0	4	0	0	0	0	278	4	0	1	0	0	5


4

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

Jurrian K. de Kanter[1,2,*], Flavia Peci[1,2,*], Eline Bertrums[1,2,3], Axel Rosendahl Huber[1,2], Anaïs van Leeuwen[1,2], Markus J. van Roosmalen[1,2], Freek Manders[1,2], Mark Verheul[1,2], Rurika Oka[1,2], Arianne M. Brandsma[1,2], Marc Bierings[1,4], Mirjam Belderbos[1,2,*], and 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
[3] Department of Pediatric Oncology/Hematology, Erasmus Medical Center, Rotterdam 3015 GD, the Netherlands
[4] Paediatric Blood and Marrow Transplant Program, University Medical Center Utrecht, Utrecht, Netherlands

Cell Stem Cell, October 7;28:1726-1739
doi:10.1016/j.stem.2021.07.012 (2021)

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.

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[1,2]. 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[3–7]. 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[8]. 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[8–10]. 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[11–13]. 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[12–16]. 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[17]. 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[18–20]. 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 (**Fig. 1A**)[21,22]. 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 (**Fig. 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[21–23]. 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 (**Fig. 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, **Fig. 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 (**Fig. S1B**). We did not observe any acquired structural variations in pre- and post-HSCT clones.

Sample	Primary diagnosis	Trans-plantation	(second) cancer	Viral reacti-vations	Antiviral therapy	(Second) cancer driver mutations C>ApA	Ref.
11396 – Dx2 AML	ALL	HSCT	AML	CMV	GCV, FC		N/A
633734 – relapse	AML	HSCT	AML-re-lapse	CMV	GCV	NRAS p.Q61K	53
103342 – relapse	AML	HSCT	AML-re-lapse	CMV	GCV, val-GCV		53
814916 – relapse	AML	HSCT	AML-re-lapse	CMV	GCV		53
AML_015	AML	HSCT	AML-re-lapse	Un-known	Unknown		52
Gondek1 – DCL	AML	HSCT	Donor cell leukemia	Un-known	Unknown	SETBP1 p.T873K	44
CPCT02090030T	Renal in-sufficiency	Kidney Tx	Vulvar carcinoma metastasis	Un-known	Unknown	HRAS, p.Q61K	50
CPCT02110076T	Renal in-sufficiency	Kidney Tx	Breast carcinoma metastasis	CMV	Val-GCV		50
CPCT02340067T	Melanoma	None	Melano-ma-relapse metastasis	None	None		50

Table 1. Clinical information of SBSA-positive cancers.
Abbreviations: ALL: acute lymphoblastic leukemia; AML: Acute myeloid leukemia; HSCT: Hematopoietic stem cell transplantation; Tx: transplantation; CMV: cytomegalovirus

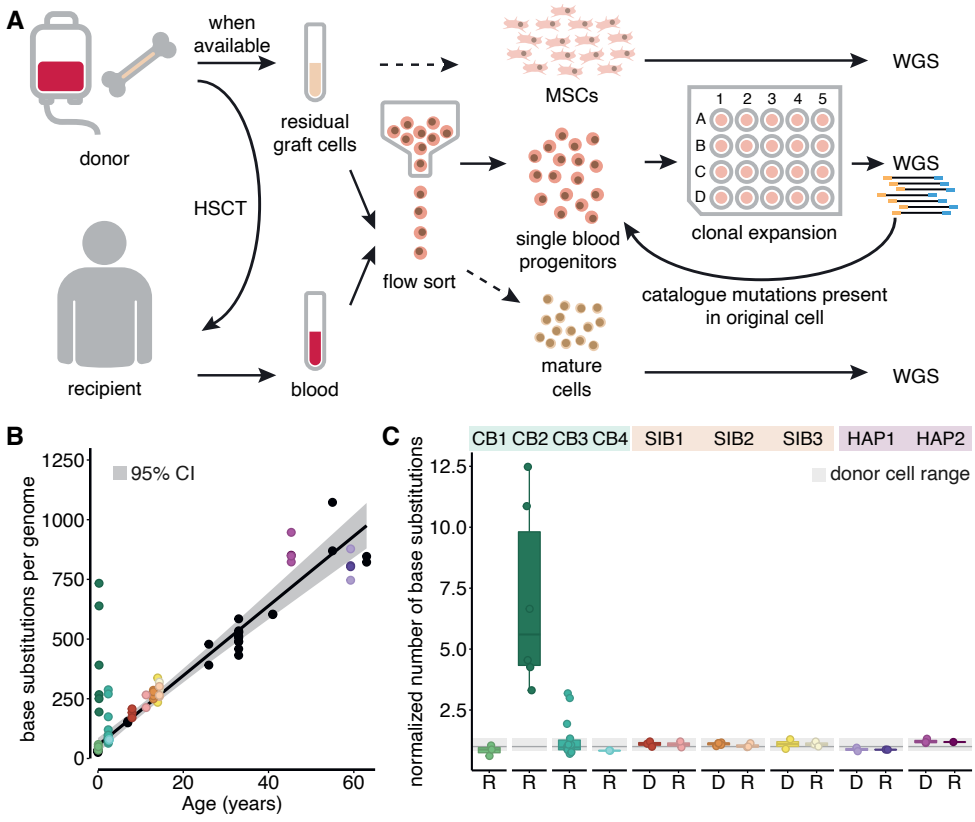


Figure 1. Mutation accumulation associated with HSCT in humans
A) Schematic representation of the experimental setup to determine somatic mutations in blood progenitor cells of hematopoietic stem cell transplantation (HSCT) donors and recipients. **B)** Correlation between the age and the number of base substitutions per genome in 32 single HSPC clones of 3 HSCT donors and 6 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% confidence interval of the baseline is depicted in gray. HSCT clones are colored similar to C, non-HSCT clones of the baseline are 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 is a single HSPC clone. In light gray, the range of the normalized number of base substitutions of donor HSPC clones is depicted. Abbreviations; CB: Cord blood; SIB: Sibling. D: HSCT donor; R: HSCT recipient. See also Figure S1 and Table 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[22]. 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 (**Fig. 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, **Fig. 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; **Fig. 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 (**Fig. 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 (**Fig. 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 (**Fig. 2**). Such signatures reflect specific mutational processes that have been active during the life of the assessed HSPCs[24–26]. 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 (**Fig. 2A-C**)[22,27,28]. This signature reflects clock-like activity of the predominant mutational process in postnatal HSPCs during healthy life[29], 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 (**Fig. 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, $p<10\text{-}e5$, **Fig. 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 (**Fig. 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 (**Fig. 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

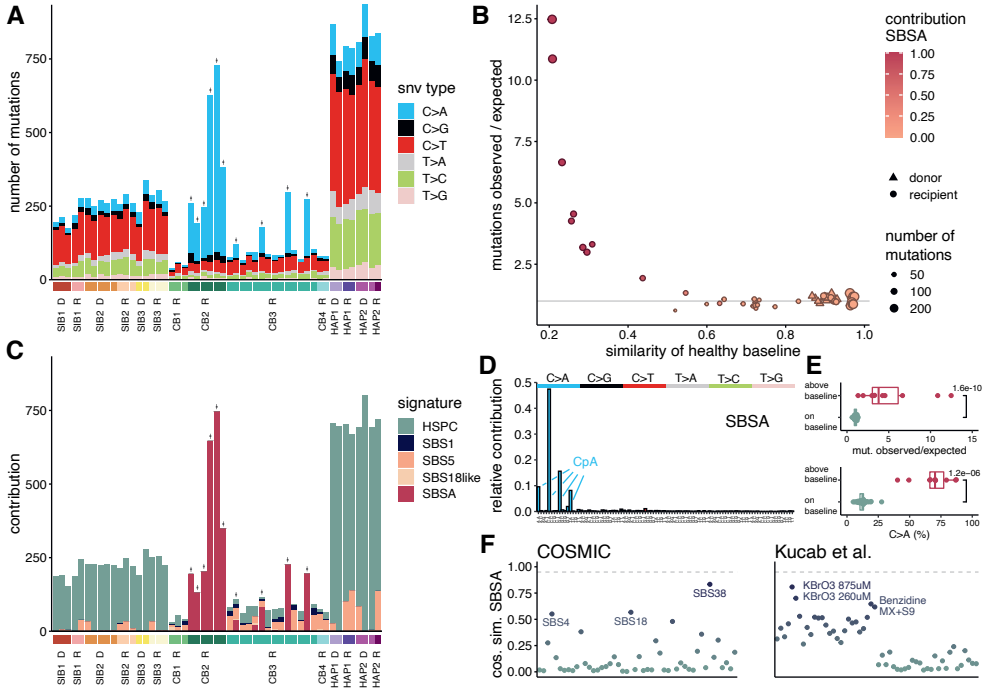


Figure 2: Transplantation-associated mutagenesis can be attributed to a unique mutational signature SBSA

A) Single base substitution (SBS) mutational spectra from HSCT donor and recipient HSPCs. "†" symbols indicate the 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 to 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 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 ageline (above, Wicoxon test). The percentage of mutations that are C>A transversion of the same groups of clones (below, Wilcoxon). **F)** The cosine similarity between the SBSA signature and SBS mutational signatures from the Cosmic v3.0 database and in vitro established signatures of environmental agents[30].

"SBSA" (**Fig. 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 (**Fig. 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

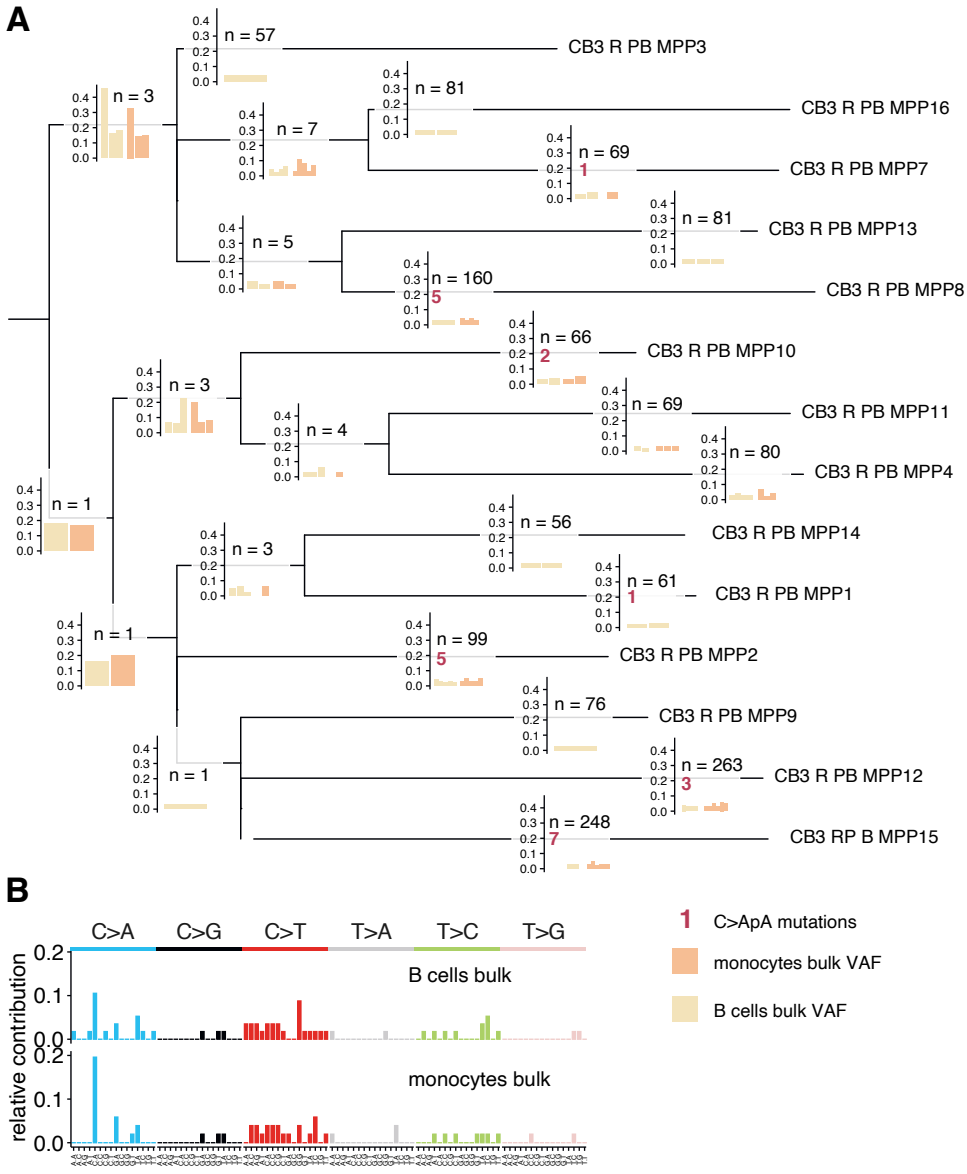


Figure 3. Detection of HSPC mutations in bulk mature populations.
A) The phylogenetic tree of the HSPCs of patient CB3 is shown. At each branch, a bar graph is plotted. The number above each bar graph indicates the total number of mutations in that branch. Each bar represents the VAF of a mutation in 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 the patients, see Figure S3.

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 (**Fig. 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 (**Fig. 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[30,31]. A cosine similarity of ≥0.95 was used to indicate that two patterns are similar[32]. We found that SBSA did not match any of the previously defined mutation signatures (**Fig. 2F**). SBSA showed highest cosine similarity with, but was still distinct from, SBS38, SBS18 and a potassium bromate (KBrO3)-induced signature (cosine similarity of 0.83, 0.57 and 0.81, respectively; **Fig. 2F, 4A and S4C**).

Molecular characterization of SBSA

SBS38, SBS18 and the KBrO3 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[24,30,33]. 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[34], 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 KBrO330 and a knockout of OGG1 (OGG1KO), which has a central role in 8-oxo-guanine base excision repair[35] (**Fig. 4A, 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 (**Fig. 4B, S4A**). In contrast, this context did not occur in the KBrO3 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 KBrO3, deficient for OGG1, and in HSPCs with a normal baseline mutation load (fdr=0.17, 0.29, 0.48 respectively, **Fig. 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[36]. 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.

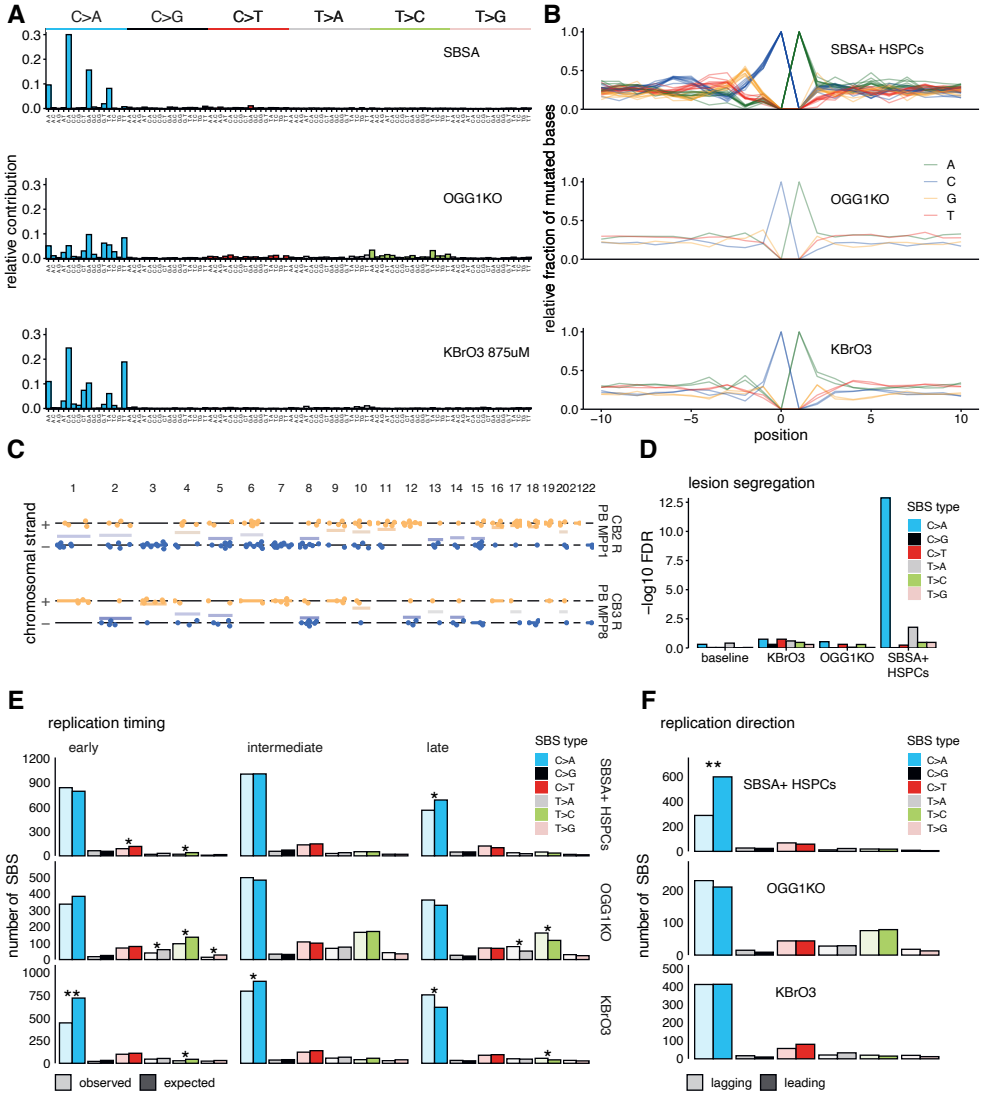


Figure 4. SBSA is characterized by lesion segregation and a strong replication direction bias.

A) SBS 96-trinucleotide mutational profiles of SBSA and oxidative stress-associated signatures of exposure to KBrO3 or knock-out of OGG1. **B)** The -10:+10 nucleotide context of C>ApA mutations of five SBSA positive HSPC clones, knock-out of OGG1 and two KBrO3-treated clones. Each line represents the mutation context in a single clone. **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, knock-out of OGG1 and exposure to KBrO3 in early, intermediate and late replicating regions. * = FDR < 0.05. **F)** Replication strand bias of the same data as depicted in E. ** = FDR < 10^{-7}. See also Figure S4.

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 (**Fig. S4B,F**)[37,38]. SBSA mutations were slightly depleted in late replicating regions of the DNA (fdr < 10e-4, **Fig. 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, **Fig. 4F, S4D**), which indicates that the mutagenic process underlying SBSA is directly coupled to DNA replication[37,38]. 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[39]. FC is a pyrophosphate analogue, which is thought to directly inhibit viral polymerase activity[40]. 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 (**Fig. 5A**). While GCV caused dose-dependent cell death at micromolar concentrations, which are also observed in human plasma (IC50 4,64 μM)[41], FC did not induce cell death at any of the tested concentrations (**Fig. 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)[42]. Both GCV and the combination treatment caused substantial DNA damage, visualized by γ-H2AX staining, while FC exposure alone did not cause considerable cell death (**Fig. 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 (**Fig. 5E**). The number of indels was similar between GCV, FC and control-treated organoids, no copy number variations or structural rearrangements were found (**Fig. 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, **Fig. 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 (**Fig. S5D-H**). Altogether, these data clearly demonstrate that GCV is the cause of the SBSA mutations.

SBSA mutations in cancer

Accumulation of somatic mutations is a key mechanism promoting carcinogenesis. To assess whether SBSA mutations can contribute to cancer development, we determined its presence in the genomes of allogeneic and autologous HSCT donors and recipients[43–48] (**Fig. 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 (**Fig. S6A, B, G**). We trained the RF on the pre- and post-HSCT HSPCs and on the healthy baseline HSPCs depicted in **Fig. 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% (**Fig. 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[27] and on a dataset of clonal hematopoiesis of indeterminate potential (CHIP) mutations in bulk WGS of 97,691 healthy individuals (**Fig. 6C**)[49]. As expected, the RF identified <1% SBSA-positive mutations in both datasets, confirming the specificity of this classifier.

Next, we applied this RF classifier to sequencing datasets of human metastatic cancers (n=3668)[50] and of hematologic disorders after allogeneic and autologous HSCT, such as clonal hematopoiesis (n=290)[43,45,47,48],therapy-related neoplasms (n=9)[44,51], and relapsed acute myeloid leukemia (AML) after allogeneic HSCT or chemotherapy(n=44)[52,53]. In total, the RF classified nine cancers of nine individual patients as SBSA-positive (**Fig. 6, Table 1**). The first was a therapy-related AML (tAML, PMC11396), in which SBSA had an estimated contribution of 28% (**Fig. 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 (**Fig. 6A**). This finding was confirmed using mutational signature analysis (**Fig. 6B**), the +/-10 nucleotide context (**Fig. S6C**) and replication strand bias (**Fig. S6H**). The C>A mutations did, however, not display a Watson-versus-Crick bias (**Fig. S6K**). Notably, although five mutations were shared between the tAML and one of the healthy HSPCs collected prior to transplantation (**Fig. 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.

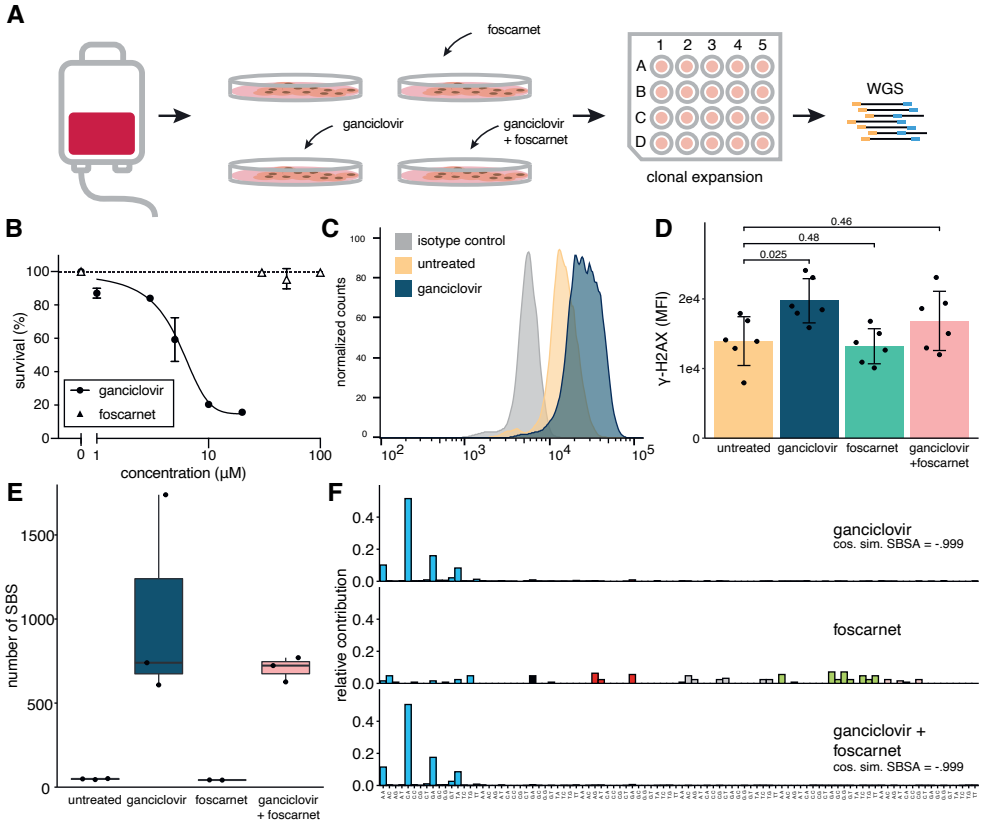


Figure 5. Ganciclovir induces SBSA mutations in vitro

A) Experimental setup of in vitro treatment of CD34+ human umbilical cord blood cells with antiviral agents Foscarnet, Ganciclovir and a combination of both. After 24 hours of treatment, single clones are sorted into 96 well plates, expanded and whole genome sequenced. **B)** Survival curve and ganciclovir treatment. For foscarnet, no curve could be fitted due to the low percentage of cell death. 200uM foscarnet is not shown and caused 86% survival. **C)** Representative histogram of γ-H2AX intensity of isotype, untreated and ganciclovir treated cord blood cells. **D)** The γ-H2AX mean fluorescence intensity (MFI) of three cord blood samples, each treated with each condition twice (Wicoxon test). See Figure S6C for values per sample, and a positive radiation control. **E)** The number of single base substitutions of each of the treatment conditions (5μM ganciclovir and/or 200μM foscarnet). **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.

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[44](**Fig. 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 (**Fig. 6C**) and had a clear SBSA signature (**Fig. 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 (**Fig. 5C,D**)[53]. Again, we confirmed this with mutational signature analysis, replication direction bias and the extended context (**Fig. S6E,J**). Also, in this case, the C>A mutations did not have a Watson-versus-Crick asymmetry (**Fig. 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(**Fig. 6E**)[50]. 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 (**Fig. 6F, S6D,I,K,L**). In contrast, the tumor of the non-transplanted melanoma patient did not show this context nor bias (**Fig. 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[54]. 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 (**Fig. 6G**). These calculations showed an increased potential of SBSA to cause stop-gain mutations (ratio of 1.6 for stop-gain compared to background) (**Fig. 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[55]. 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-

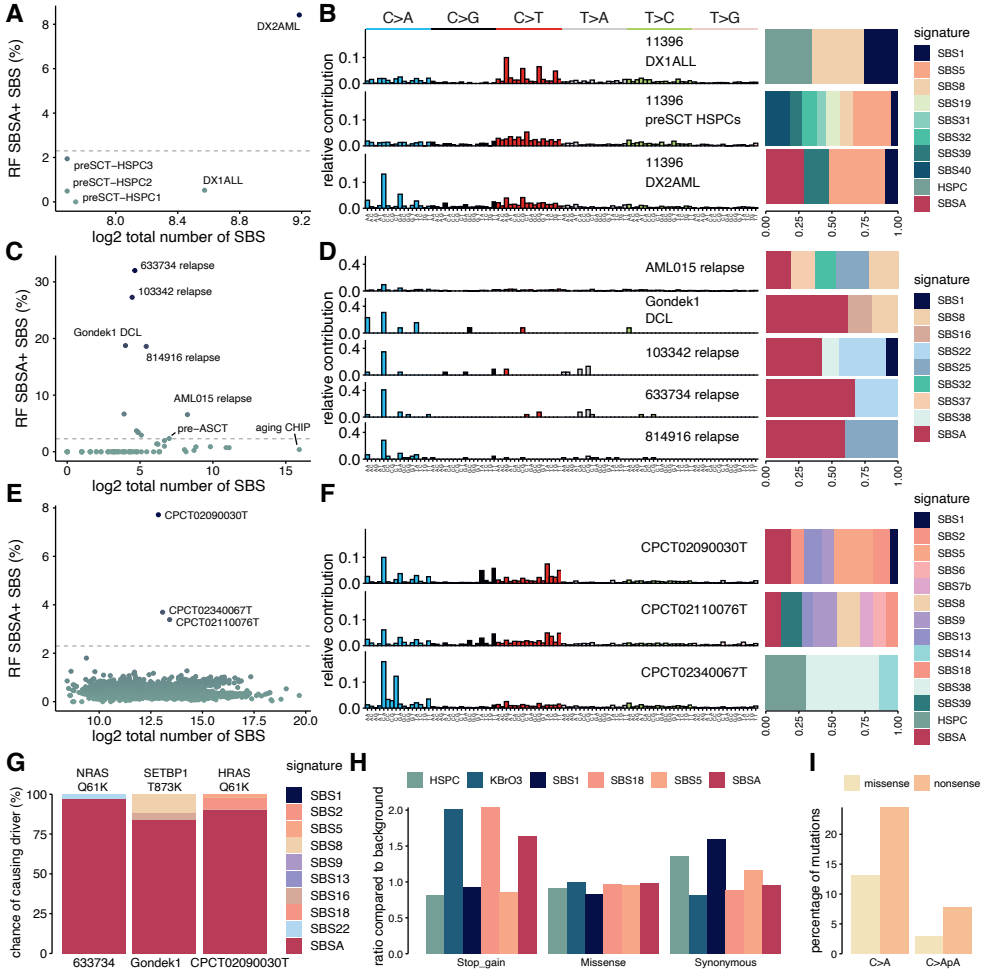


Figure 6. SBSA is present in transplant-related cancers and can cause cancer driver mutations.

The percentage of random forest-predicted SBSA mutations compared to the total number of mutations in samples of **(A)** patient PMC11396, **(C)** targeted and WGS mutation datasets of autologous and allogeneic SCT grafts and recipients, normal aging, age-associated CHIP, post-HSCT AML relapses and post-HSCT tMN cases, **(E)** a Dutch WGS cohort of 3668 solid tumor metastases[50]. 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 therapy-related AML of patient PMC11396. **D)** Similar to B, but of the SBSA positive samples from Cl44. DCL = donor cell leukemia. **F)** Similar to B, but of metastases that are SBSA positive predicted by the random forest in a Dutch cohort of 3668 solid tumor metastases[50]. **G)** Probability estimation of each signature in a tumor causing C>ApA driver mutations. **H)** The potential mutational impact of six SBS mutational signatures, including SBSA, in blood cancer driver genes, normalized to a "flat" background signatures 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.

context, confirming our previous results (**Fig. 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[27,29]. As HSCs divide approximately every 40 weeks[56], 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[27]. 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[57,58]. 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[59–61]. 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[22,27,62].

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[63]. However,

antiviral nucleoside analogues have also been reported to mediate their effect by inducing lethal mutagenesis of the viral genome[64]. 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[38]. 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[65]. 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[27], 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.

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.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-yH2A.X (Ser139) FITC (Merk) or Mouse IgG isotype antibody were added to samples and cells were incubated for 20 min on ice. After staining, cells were washed with 0.1% saponin in PBS and resuspended in FACS buffer (1x PBS, 2-5% FBS, 2mM EDTA, 2mM NaN_3) prior to flow cytometric analysis. For analysis of single-cell mutagenesis caused by antiviral treatment, CD34+ cells were sorted as single cells into flat-bottom 384-

well plates (Greiner), using the same antibody mix and sorting strategy as for bone marrow and peripheral blood HSPCs. Cells were clonally expanded for 4-6 weeks, after which DNA was isolated (QIAamp DNA micro kit, Qiagen) and sent for whole genome sequencing.

Analysis of γ-H2AX expression by flow cytometry.
After drugs incubation, cells were harvested and washed with PBS. 100.000-200.000 CD34+ cells were resuspended in ice-cold fixative solution (2.5% formaldehyde and 0.93% methanol in sterile filtered PBS), incubated for 20 min at 4°C and transferred to a 96 well plate. Fixed samples were washed twice with PBS. Next, cells were resuspended in permeabilization buffer containing 0.5% saponin, 0.5% BSA, 10mM HEPES, 140mM NaCl, 2.5mM CaCl2 in water, pH 7.4, sterile filtered. Anti-yH2A.X (Ser139) FITC (Merk) or Mouse IgG isotype antibody 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' [66]. 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 [67]. 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[68]. All resulting structural variants were checked by hand in the IGV[69] 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[21,70]. 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 **Fig. 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[71]. The 0.95 confidence interval was calculated using the ggeffects package in R [72]. 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)[32] 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[22,32]. The five extracted mutational patterns were compared to the COSMIC v3 signatures[31] together with our previously identified HSPC signature[22] 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[31] and signatures from Kucab et al[30]) 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[36], 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[73].

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[30]. 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[74].

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[75]. 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 [76] 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"[77]. (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)[78]. (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)[79]. (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[80].

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[81]. The normal aging dataset used as control for the RF was extracted from the supplementary table of Lee-Six et al.[27]. The AML relapse data were obtained from Christopher et al.[53] and Stratmann et al.[52]. Data on the post-HSCT neoplasms were obtained from Berger et al.[51] and Gondek et al[44]. 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[80]. The aging CHIP dataset was obtained from Bick et al[49].

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[82].

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).

References

1. Pasquini, M. C., Wang, Z., Horowitz, M. M. & Gale, R. P. 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 (2010).
2. Passweg, J. R. et al. Hematopoietic stem cell transplantation in Europe 2014: more than 40 000 transplants annually. Bone Marrow Transplant. 51, 786–792 (2016).
3. Aiuti, A. et al. Lentiviral hematopoietic stem cell gene therapy in patients with wiskott-aldrich syndrome. Science 341, 1233151 (2013).
4. Aiuti, A. et al. Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. Science 296, 2410–2413 (2002).
5. Dunbar, C. E. et al. Gene therapy comes of age. Science (80-.). 359, 1–10 (2018).
6. De Ravin, S. S. et al. Lentiviral hematopoietic stem cell gene therapy for X-linked severe combined immunodeficiency. Sci. Transl. Med. 8, 1–11 (2016).
7. Xu, L. et al. CRISPR-Edited Stem Cells in a Pateint with HIV and Acute Lymphocytic Leukemia. N. Engl. J. Med. 381, 1240–1247 (2019).
8. Bhatia, S. Long-term health impacts of hematopoietic stem cell transplantation inform recommendations for follow-up. Expert Rev. Hematol. 4, 437–454 (2011).
9. Clark, C., Savani, M., Mohty, M. & Savani, B. What do we need to know about allogeneic hematopoietic stem cell transplant survivors? Bone Marrow Transplantation2 51, 1025–1031 (2016).
10. Majhail, N. S. et al. Prevalence of Hematopoietic Cell Transplant Survivors in the United States. Biol. Blood Marrow Transplantation2 19, 1498–1501 (2013).
11. Kuijk, E. et al. The mutational impact of culturing human pluripotent and adult stem cells. Nat. Commun. 11, 1–12 (2020).
12. Thompson, O. et al. Low rates of mutation in clinical grade human pluripotent stem cells under different culture conditions. Nat. Commun. 11, 1–14 (2020).
13. Yamanaka, S. Pluripotent stem cell-based therapy - Promise and challenges. Cell Stem Cell 27, 523–531 (2020).
14. Andrews, P. W. et al. Assessing the Safety of Human Pluripotent Stem Cells and Their Derivatives for Clinical Applications. Stem Cell Reports 9, 1–4 (2017).
15. Avior, Y., Eggan, K. & Benvenisty, N. Cancer-related mutations identified in primed and naive pluripotent stem cells. Cell Stem Cell 25, 456–461 (2019).
16. Lamm, N. et al. Genomic Instability in Human Pluripotent Stem Cells Arises from Replicative Stress and Chromosome Condensation Defects. Cell Stem Cell 18, 253–261 (2016).
17. Mandai, M. et al. Autologous Induced Stem-Cell–Derived Retinal Cells for Macular Degeneration. N. Engl. J. Med. 376, 1038–1046 (2017).
18. Collins, F. S. & Gottlieb, S. The next phase of human gene-therapy oversight. New England Journal of Medicine vol. 379 1393–1395 at https://doi.org/10.1056/NEJMp1810628 (2018).
19. Hacein-Bey-Abina, S. et al. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J. Clin. Invest. 118, 3132–3142 (2008).
20. Howe, S. J. et al. Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J. Clin. Invest. 118, 3143–3150 (2008).
21. Jager, M. et al. Measuring mutation accumulation in single human adult stem cells by whole-genome sequencing of organoid cultures. Nat. Protoc. 13, 59–78 (2017).
22. Osorio, F. G. et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. 25, 2308–2316 (2018).
23. Rosendahl Huber, A., Manders, F., Oka, R. & van Boxtel, R. Characterizing Mutational Load and Clonal Composition of Human Blood. JoVE (2019) doi:10.3791/59846.
24. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
25. Alexandrov, L. B. et al. Mutational signatures associated with tobacco smoking in human cancer. Science (80-.). 354, 618–622 (2016).
26. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature 513, 422–5 (2014).
27. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. Nature 561, 473–478 (2018).
28. Maura, F. et al. A practical guide for mutational signature analysis in hematological malignancies. Nat. Commun. 10, 1–12 (2019).
29. Hasaart, K. et al. Mutation accumulation and developmental lineages in normal and Down syndrome human fetal haematopoiesis. Sci. Rep. 10, 12991 (2020).

30. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. Cell 177, 821–836 (2019).
31. Tate, J. G. et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 47, D941–D947 (2019).
32. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 10, 1–11 (2018).
33. Brem, R., Macpherson, P., Guven, M. & Karran, P. 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 (2017).
34. Pleguezuelos-Manzano, C. et al. Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. Nature 580, 269–273 (2020).
35. Boiteux, S., Coste, F. & Castaing, B. 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 Radical Biology and Medicine vol. 107 179–201 at https://doi.org/10.1016/j.freeradbiomed.2016.11.042 (2017).
36. Aitken, S. S. J. et al. Pervasive lesion segregation shapes cancer genome evolution. Nature 583, 265–270 (2020).
37. Haradhvala, N. J. et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. Cell 164, 538–549 (2016).
38. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol. 19, 129 (2018).
39. 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 Research vol. 154 66–86 at https://doi.org/10.1016/j.antiviral.2018.04.004 (2018).
40. Crumpacker, C. S. Mechanism of action of foscarnet against viral polymerases. Am. J. Med. 92, S3–S7 (1992).
41. Piketty, C. et al. Monitoring plasma levels of ganciclovir in AIDS patients receiving oral ganciclovir as maintenance therapy for CMV retinitis. Clin. Microbiol. Infect. 6, 117–120 (2000).
42. Maggs, D. J. & Clarke, H. E. In vitro efficacy of ganciclovir, cidofovir, penciclovir, foscarnet, idoxuridine, and acyclovir against feline herpesvirus type-1. Am. J. Vet. Res. 65, 399–403 (2004).
43. Boettcher, S. et al. Clonal hematopoiesis in donors and long-term survivors of related allogeneic hematopoietic stem cell transplantation. Blood 135, 1548–1559 (2020).
44. Gondek, L. P. et al. Donor cell leukemia arising from clonal hematopoiesis after bone marrow transplantation. Leukemia vol. 30 1916–1920 at https://doi.org/10.1038/leu.2016.63 (2016).
45. Husby, S. et al. Clinical impact of clonal hematopoiesis in patients with lymphoma undergoing ASCT: a national population-based cohort study. Leukemia 34, 3256–3268 (2020).
46. Lombard, D. B. et al. DNA repair, genome stability, and aging. Cell 120, 497–512 (2005).
47. Mouhieddine, T. H. et al. Clonal hematopoiesis is associated with adverse outcomes in multiple myeloma patients undergoing transplant. Nat. Commun. 11, 1–9 (2020).
48. Ortmann, C. A. et al. Functional Dominance of CHIP-Mutated Hematopoietic Stem Cells in Patients Undergoing Autologous Transplantation. Cell Rep. 27, 2022–2028 (2019).
49. Bick, A. G. et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature 586, 763–768 (2020).
50. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumors. Nature 575, 210–216 (2019).
51. Berger, G. et al. Early detection and evolution of preleukemic clones in therapy-related myeloid neoplasms following autologous SCT. Blood 131, 1846–1857 (2018).
52. Stratmann, S. et al. Genomic characterization of relapsed acute myeloid leukemia reveals novel putative therapeutic targets. Blood Adv. 5, 900–912 (2021).
53. Christopher, M. J. et al. Immune Escape of Relapsed AML Cells after Allogeneic Transplantation. N. Engl. J. Med. 379, 2330–2341 (2018).
54. Morganella, S. et al. The topography of mutational processes in breast cancer genomes. Nat. Commun. 7, 1–11 (2016).
55. Jaiswal, S. et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. N. Engl. J. Med. 371, 2488–2498 (2014).
56. Caitlin, S. N. et al. The replication rate of human hematopoietic stem cells in vivo. Blood 117, 4460–4466 (2011).
57. Biasco, L. et al. In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. Cell Stem Cell 19, 107–119 (2016).
58. Scala, S. et al. Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. Nat. Med. 24, 1683–1690 (2018).
59. Weissman, I. L. Stem cells: Units of development, units of regeneration, and units in evolution. Cell vol. 100 157–168 at https://doi.org/10.1016/S0092-8674(00)81692-X (2000).
60. Sun, J. et al. Clonal dynamics of native haematopoiesis. Nature 514, 322–7 (2014).

61. Lu, R., Czechowicz, A., Seita, J., Jiang, D. & Weissman, I. L. Clonal-level lineage commitment pathways of hematopoietic stem cells in vivo. Proc. Natl. Acad. Sci. U. S. A. 116, 1447–1456 (2019).
62. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. Nature 593, 405–410 (2021).
63. Chen, H., Beardsley, G. P. & Coen, D. M. 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 (2014).
64. Loeb, L. A. et al. Lethal mutagenesis of HIV with mutagenic nucleoside analogs. Proc. Natl. Acad. Sci. U. S. A. 96, 1492–1497 (1999).
65. Griffiths, P. & Lumley, S. Cytomegalovirus. Current Opinion in Infectious Diseases vol. 27 554–559 at https://doi.org/10.1097/QCO.0000000000000107 (2014).
66. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–9 (2009).
67. Depristo, M. A. et al. A framework for variation discovery and genotyping using nextgeneration DNA sequencing data. Nat. Genet. 43, 491–498 (2011).
68. Cameron, D. L. et al. GRIDSS, PURPLE, LYNX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number. BioRxiv (2019) doi:https://doi.org/10.1101/781013.
69. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Br. Bioinforma. 14, 178–192 (2013).
70. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature 538, 260–264 (2016).
71. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48 (2015).
72. Lüdecke, D. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. J. Open Source Softw. 3, 772 (2018).
73. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 57, 289–300 (1995).
74. Weiner, J. riverplot: Sankey or Ribbon Plots. at https://cran.r-project.org/package=riverplot (2017).
75. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773 (2019).
76. Liaw, A. & Wiener, M. Classification and Regression by randomForest. R News 2, 18–22 (2002).
77. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
78. Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): Data portal update. Nucleic Acids Res. 46, D794–D801 (2018).
79. Stunnenberg, H. G. et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell vol. 167 1145–1149 at https://doi.org/10.1016/j.cell.2016.11.007 (2016).
80. LiftOver - USCS Genome Browser.
81. Burns, S. S. & Kapur, R. Stem Cell Reports Review Clonal Hematopoiesis of Indeterminate Potential as a Novel Risk Factor for Donor-Derived Leukemia. Stem Cell Reports 15, 279–291 (2020).
82. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York, 2016).

Supplementary Material

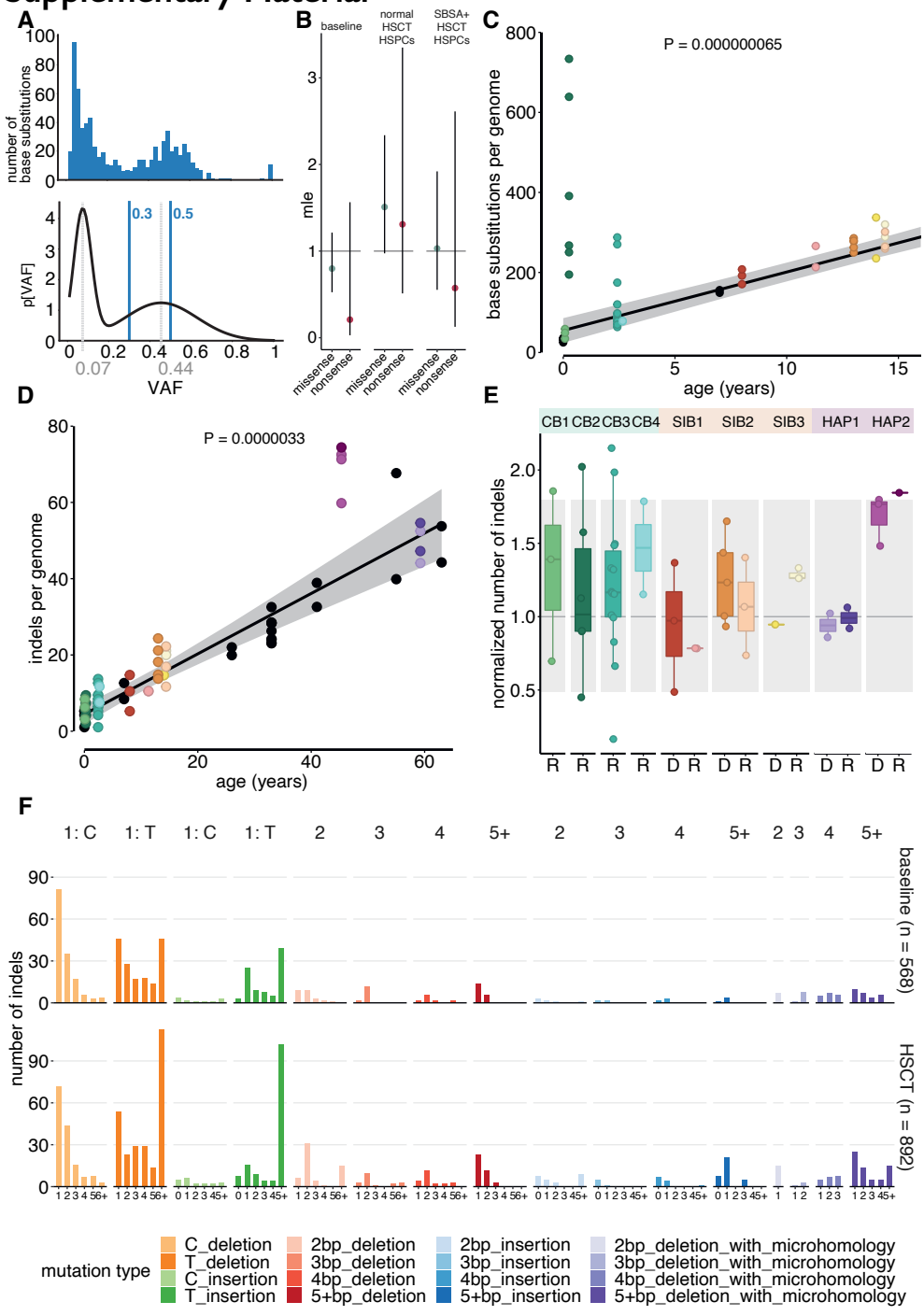


Figure S1. The number of indels in HSCT donor and recipient HSPCs is variable but not consistently altered after transplantation, related to Figure 1. *For the legend, see page 130.*

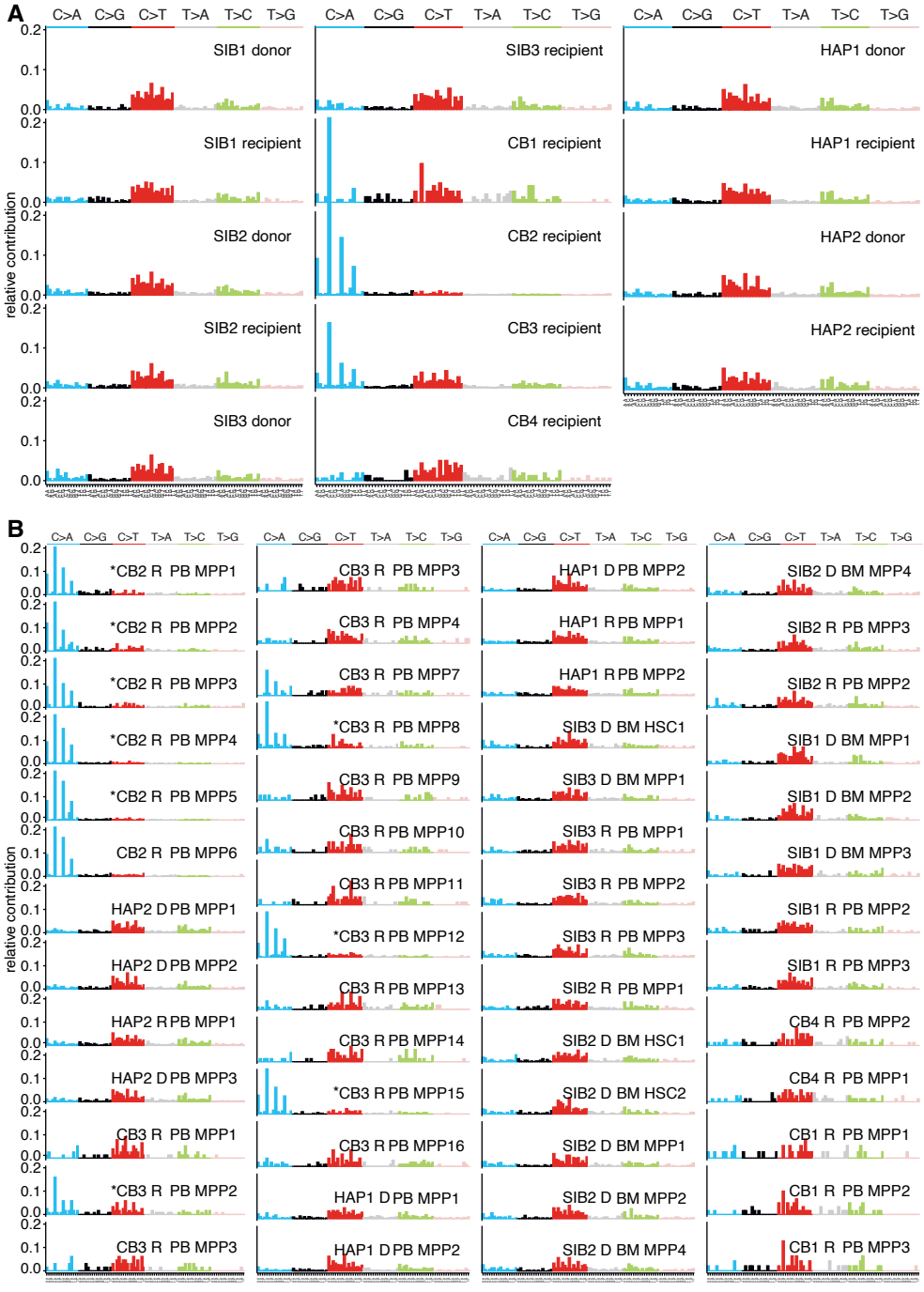


Figure S2. Verification of WGS results by phylogenetic analyses and re-sequencing, related to Figure 2. *For the legend, see page 130.*

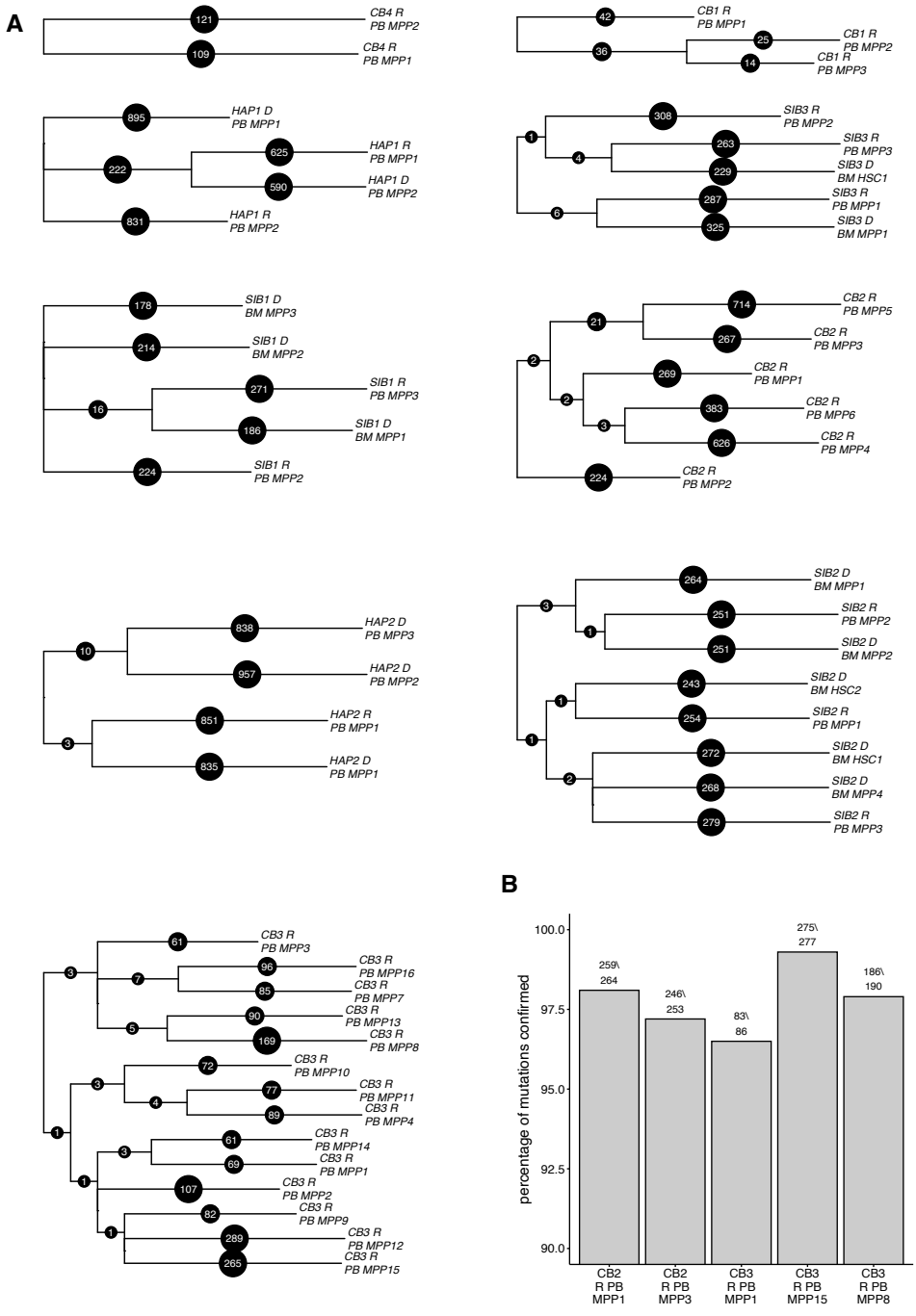


Figure S3. HSCT recipient clones with increased mutational burden have higher contribution of NpC>ApA mutations, related to Figure 3.

(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 "*".

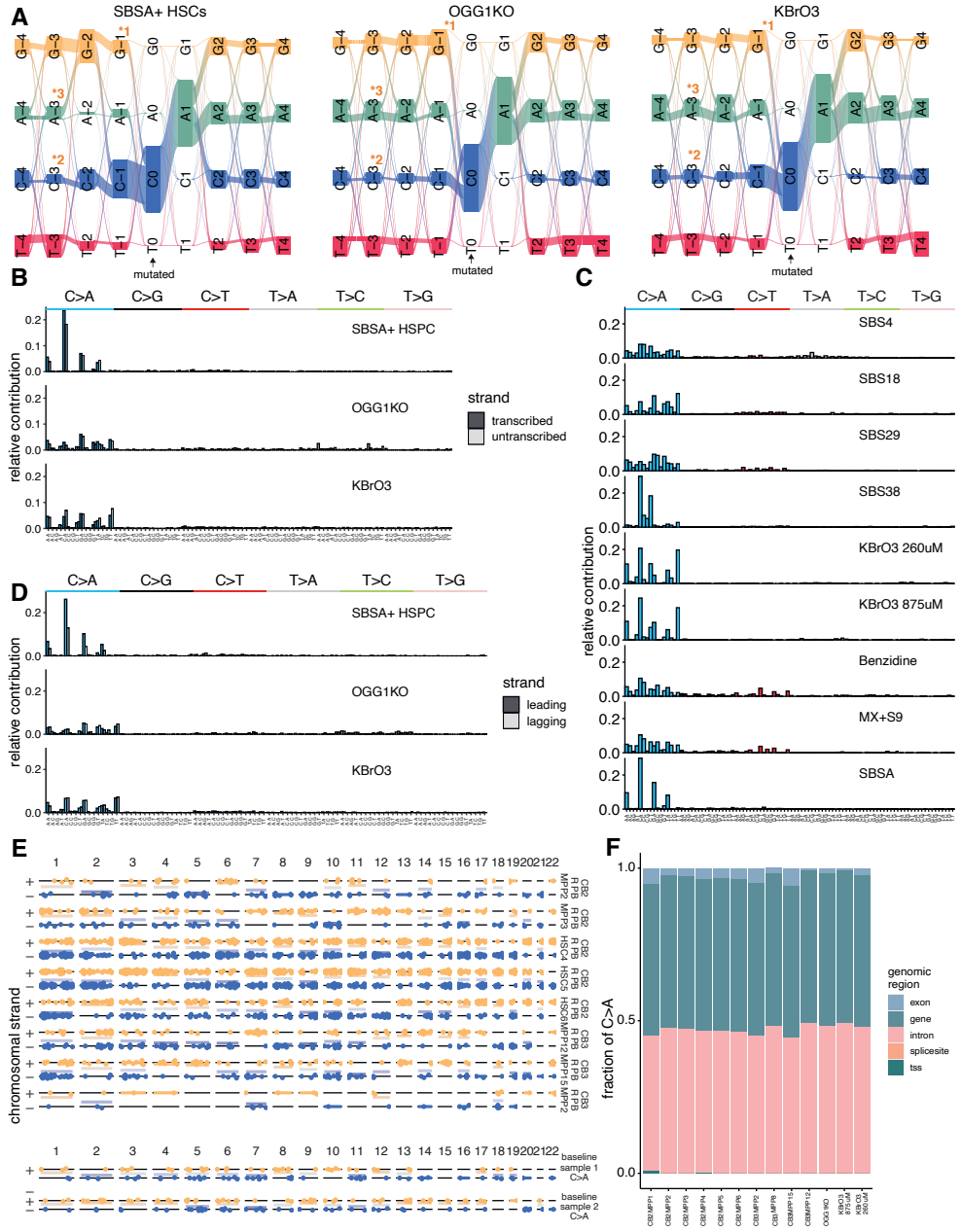


Figure S4. SBSA has a distinct nucleotide context and other characteristics that are distinct from previously reported signatures, related to Figure 4. *For the legend, see page 130.*

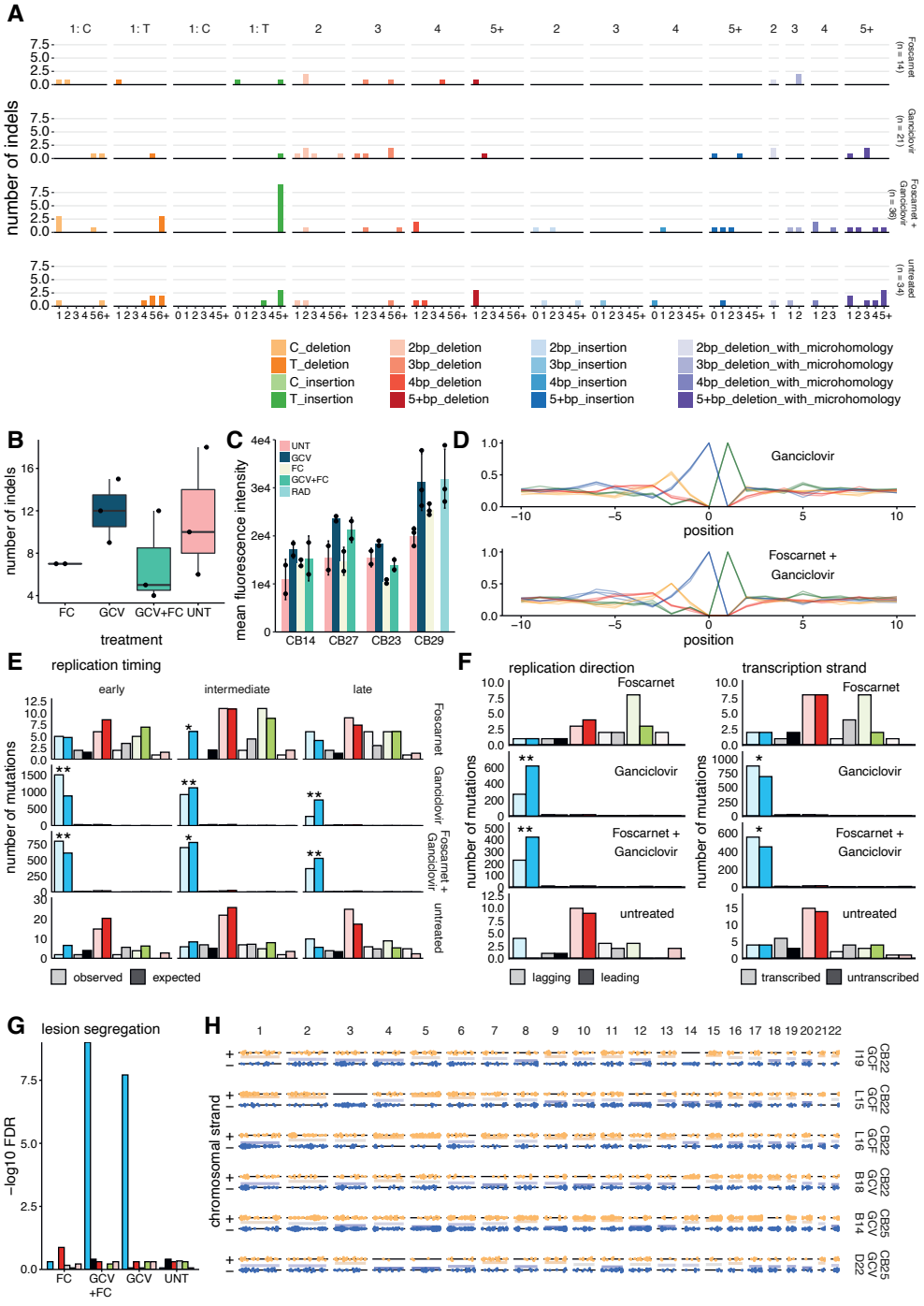


Figure S5. Mutations induced in vitro by ganciclovir have similar molecular characteristics as SBSA, related to Figure 5. *For the legend, see page 130.*

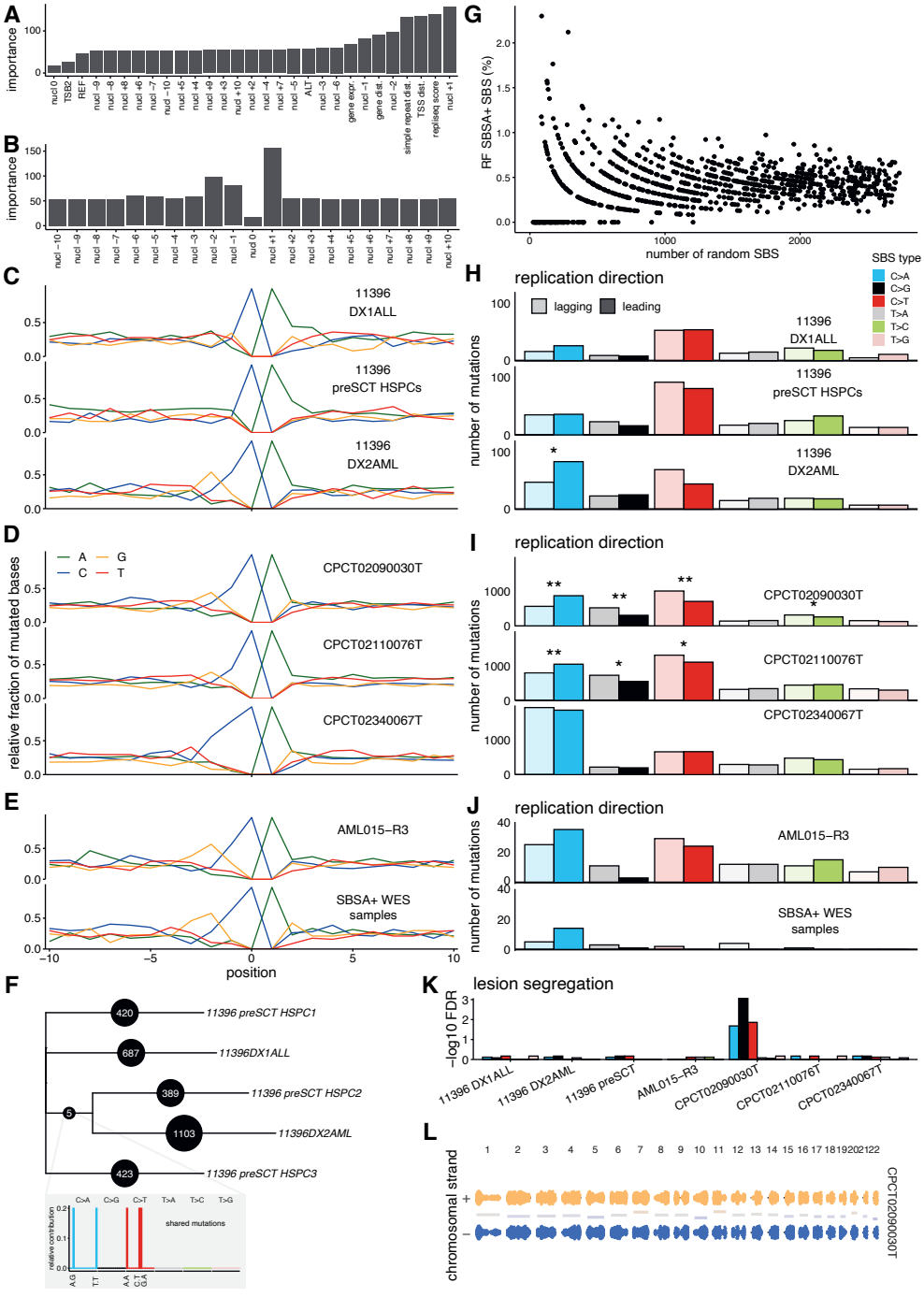


Figure S6. A random forest-based approach identifies tumors with contribution of SBSA to their mutational profile, related to Figure 6. *For the legend, see page 131.*

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 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 ITEMS

Supplementary Table S1. An overview of the clinical data.
Available online (QR code below)

Supplementary Table S2. An overview of the samples and the number of mutations.
Available online (QR code below)

Supplementary Table S3. The somatic base substitutions and indels called in HSCT recipient and donor HSPC clones.
Available online (QR code below)

Supplementary Table S4. The 96-trinucleotide mutation profile of SBSA.
Available online (QR code below)


5

The genomic safety of antiviral nucleoside analogs in hematopoietic stem cells

Jurrian K. de Kanter[1,2*], Flavia Peci[1,2*], Niels Groenen[1,2], Laurianne Trabut[1,2], Lucca Derks[1,2], Freek Manders[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

Nucleoside analogs (NAs) are widely used in the treatment of cancer, fungal infections, and viral infections. NAs can not only be incorporated in the DNA of infected cells, but also in the DNA of uninfected cells, although at a lower rate. However, the exact mutational consequences of NA exposure in uninfected cells remain unclear. Recently, we showed that the antiviral NA ganciclovir induces a high number of mutations in healthy hematopoietic stem and progenitor cells (HSPCs). To test if this mutagenesis is characteristic for all NAs, we systematically assessed the mutations induced by fourteen antiviral NAs using whole genome sequencing of in vitro exposed human HSPCs. The majority of NAs did not induce mutations. However, treatment with five clinically approved antiviral NAs resulted in significantly more base substitutions compared to untreated conditions. The NA structures and mutational profiles suggested that most NAs are incorporated into the genome, but that some NAs are mutagenic without being incorporated into the DNA. Finally, molnupiravir-induced mutations that we identified were markedly different from those found in SARS-CoV-2 genomes, indicating differences in genomic incorporation and repair in human and viral cells. Studying NA mutagenicity is therefore important to assess their clinical safety and for understanding the replication and DNA damage repair machinery.

Introduction

The importance of vaccinations in preventing symptomatic viral infections has been underlined by the recent SARS-CoV-2 pandemic. However, other treatments are needed for infections in immunocompromised individuals, for viruses without an available vaccine, and for infections in unvaccinated patients[1–3]. Nucleoside analogs (NAs) are a category of drugs that are effective in the treatment of among others HIV[4], herpes[5,6], and corona viruses[7–9]. They are also used to treat fungal infections and cancer[10,11]. NAs are created by synthetically modifying a naturally occurring nucleoside or nucleotide (i.e., a phosphorylated nucleoside). The nucleobase, the sugar moiety, and the glycosidic bond between them all have been altered to create NAs[12]. Antiviral NAs operate via three main pathways. First, they can act as chain terminators, making DNA/RNA chain elongation impossible upon incorporation[13]. Second, they can directly inhibit the polymerase functioning without genomic incorporation, for example by preventing replication initiation[14]. Third, NAs can render viruses nonfunctional by being incorporated in the viral genome and subsequently inducing a high number of mutations, a phenomenon called lethal mutagenesis[15].

Presumably, NAs that cause a high number of mutations can only be safely used if they are specific for virally infected cells. Nucleosides and NAs can only be incorporated into DNA/RNA in their triphosphate form. Some NAs are specifically mutagenic in infected cells, as the first phosphorylation step is much more efficiency performed

by viral kinases compared to human kinases[16,17]. Other NAs are not selectively phosphorylated but are more efficiently incorporated by viral polymerases than human polymerases[18]. Still, most NAs are phosphorylated in uninfected human cells and incorporated into the genome, just at a much lower rate than in virally infected cells[16,17]. Therefore, it is important to assess the consequences of NA treatment in uninfected cells to ensure clinical safety.

Current pre-clinical safety testing is done by among others bacterial reverse mutation tests (AMES test), mutagenicity tests based on single human genes (HPRT, XPRT, TK kinase), micronuclei tests, chromosomal analyses, and metabolomic toxicology screening[19,20]. Finally, in vivo toxicity is assessed in mice, for example by embryo-fetal development studies[21]. None of these are able to comprehensively assess genome-wide mutation numbers induced by the drug. Recent work from our group and others using in vitro treatment coupled to whole-genome sequencing (WGS) has shown that ganciclovir (GCV) treatment is mutagenic to non-infected human cells[22,23]. Here, we use the same highly sensitive, standardized method to systematically screen fourteen antiviral NAs in human HSPCs using WGS[24]. Our screen indicates that, besides ganciclovir, five out of fourteen NAs induce a significant number of mutations, and that the other nine are not mutagenic.

Results

Most antiviral NAs are not mutagenic in human CB-HSPCs.

The mutagenicity of antiviral NAs was tested by a previously published in vitro method[22,24] (**Fig. 1A**). Briefly, HSPCs derived from human umbilical cord blood (UCB) were exposed to different concentrations of the compound for 4 days. Cells treated with the IC40-IC60 concentration were clonally expanded to obtain sufficient DNA for WGS. Mutations present in the initial colony-forming cell were separated from artifacts and mutations acquired in vitro based on the mutations' variant allele frequency (VAF)[24]. Using this method, the mutagenicity of 14 FDA-approved NAs was tested. These included treatments for endemic viruses like herpes viruses and SARS-CoV-2, and drugs used to treat hepatitis and HIV (**Table 1, Table S1**). For each NA, 3 clones of the same donor were sequenced, except for EMT (two cells) and ACV (four cells from two donors, see **Methods**). As a control, we applied WGS on 20 untreated HSPC clones from 7 independent UCB donors. Finally, we included previously published WGS data of the highly mutagenic NA ganciclovir[22].

No SVs or CNVs were detected in any of the treated clones and no significant increase in the number of small insertion/deletions (indels) was detected after treatment with any NA (**Fig. 1B**). Furthermore, only the treatment with zidovudine (AZT), zalcitabine (DDC), and lamivudine (3TC) resulted in a minor increase in the number of double base substitutions (DBS) compared to the untreated clones (**Fig. 1C**). Therefore, we focused on single base substitutions (SBS, **Fig. 1D**). In untreated clones, an average of 38 SBS (CI 95%: 34-43) were present. This low

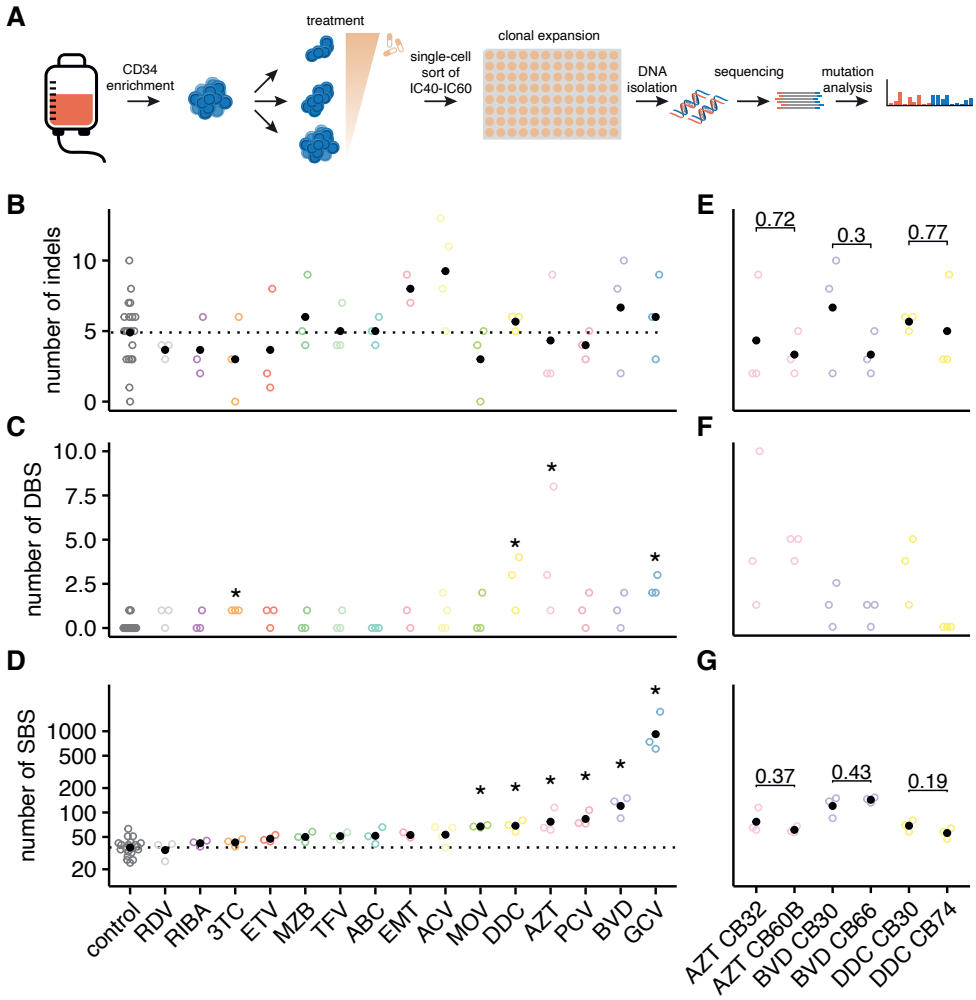


Figure 1. A subset of antiviral nucleoside analogues induces single and double base substitutions

A) Experimental setup of the screening method. **B)** The number of small insertions and deletions (indels) 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 indels found in a single clone. **C)** The same clones as in B, but the number of double base substitutions (DBS) are shown. "*" indicate treatments for which the number of DBS was significantly different from controls ($p<0.05$). P-values were calculated by the Fisher test (taking any value higher than 0 as positive) and fdr-corrected. **D)** The same clones as in C, but the number of single base substitutions (SBS) are shown. P-values were calculated by the Wilcoxon test and fdr-corrected. **E)** For the three most mutagenic NAs tested here, the treatment was repeated in cells of a second UBC donor. The number of indels are shown for the two biological duplicates. P-values are calculated using the Wilcoxon test. **F)** The same as in E, but the number of double base substitutions (DBS) are shown. **G)** the same as in E, but the number of SBS are shown.

background mutation load in UCB-HSPCs allows the identification of even lowly mutagenic drugs. Clones exposed to five out of the fourteen tested NAs harbored a significantly increased number of SBS compared to untreated cells (**Fig. 1D**, p = 0.225, fdr-corrected Wilcoxon test). These were brivudine, penciclovir, zidovudine, zalcitabine, and molnupiravir. No compound was as mutagenic as ganciclovir, which induced an average of 991 mutations per cell. The next most mutagenic compound (brivudine) induced less than one tenth of that (86 mutations). To confirm these results, the treatment of zidovudine, brivudine and zalcitabine was repeated with cells 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. 1E-G, S1B,C**).

Nucleoside analog	abbreviation	mutagenic	purine/ pyrimidine	nucleotide	sugar 3' OH	sugar 2' OH	nucleo base change	cyclic sugar	other sugar changes	glycosidic bond change	5' PO3H	treatment of viruses
Ganciclovir	GCV	Y	purine	G	Y	-	-	-	-	-	-	CMV
Acyclovir	ACV	N	purine	G	-	-	-	-	-	-	-	Herpes viruses
Penciclovir	PCV	Y	purine	G	Y	-	-	-	Y	-	-	Herpes virusses
Brivudine	BVD	Y	pyrimidine	T	Y	-	Y	Y	-	-	-	HZ
Ribavirin	RIBA	N	purine	G	Y	Y	Y	Y	-	-	-	RSV/HCV
Remdesivir	RDV	N	purine	A	Y	Y	Y	Y	Y	-	-	SARS-CoV-2/HCV
Zidovudine	AZT	Y	pyrimidine	T	-	-	-	Y	Y	-	-	HIV
Abacavir	ABC	N	purine	G	-	-	Y	Y	Y	-	-	HIV
Mizoribine	MZB	N	purine	G	Y	Y	Y	Y	-	-	-	Renal Tx
Zalcitabine	DDC	Y	pyrimidine	C	-	-	-	Y	-	-	-	HIV
Tenofovir	TFV	N	purine	G	-	-	-	-	-	Y	Y	HIV/HBV
Molnupiravir (NHC)	MOV	Y	pyrimidine	C	Y	Y	Y	Y	-	-	-	SARS-CoV-2
Entecavir	ETV	N	purine	G	Y	-	-	Y	Y	-	-	HBV
Lamivudine	3TC	N	pyrimidine	C	-	-	-	Y	Y	-	-	HIV/HBV
Emtricitabine	EMT	N	pyrimidine	C	-	-	Y	Y	Y	-	-	HIV

Table 1. Information on the nucleoside analogues tested in this manuscript.
Y = yes, N = no.

When inspecting the structure of the NAs, including ganciclovir, no common features could be identified that separated the mutagenic compounds from the non-mutagenic compounds (**Fig. S2**). Molnupiravir was the only mutagenic NA with a 2' hydroxyl group on the sugar moiety, which is normally present on the ribose of RNA molecules, but not the deoxyribose of DNA (**Table 1**). In addition, zalcitabine and zidovudine were the only mutagenic NAs without a 3' hydroxyl group, which is needed to form the phosphodiester bond with the following nucleotide during DNA elongation. Similarly, two out of six compounds had changes to the nucleobase, and

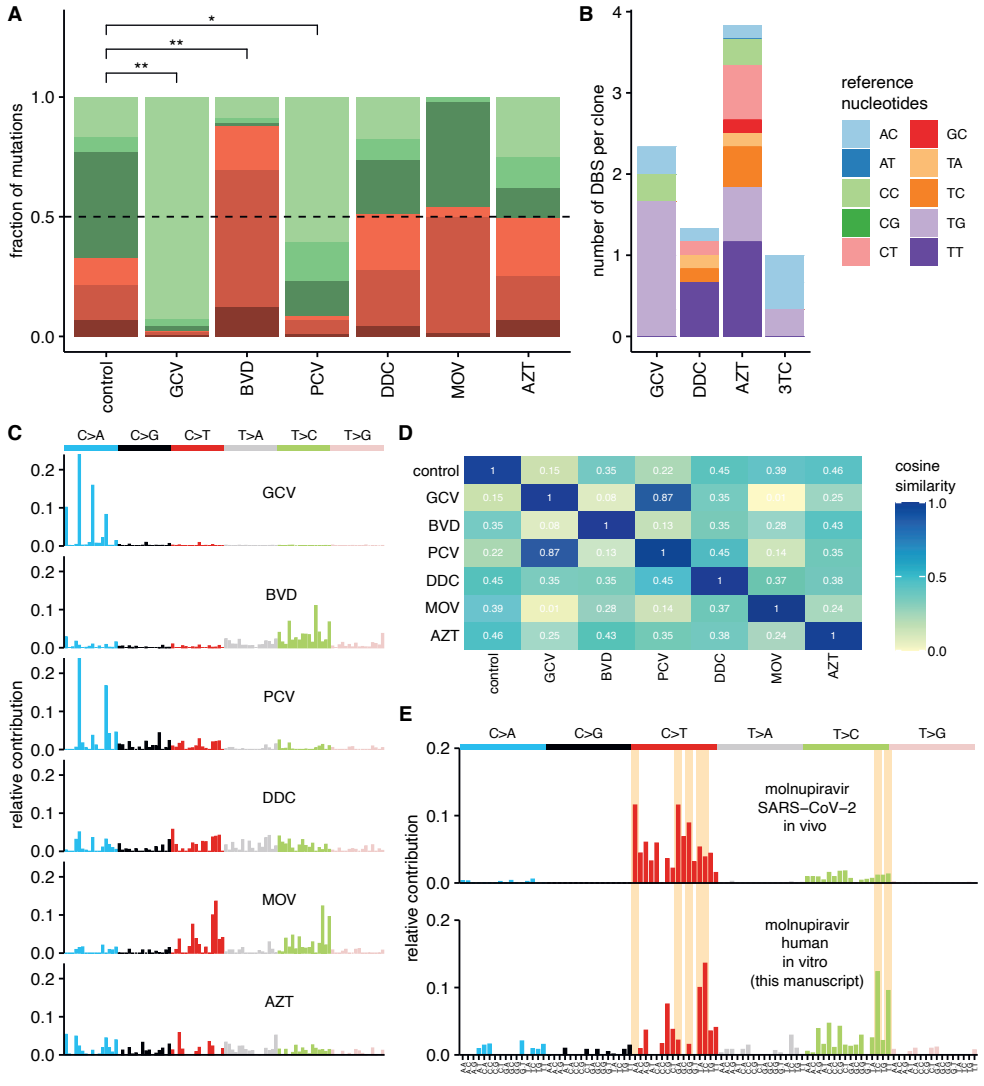


Figure 2. Each NA induces mutations in unique contexts

A) The fraction of the six single base substitution (SBS) types, as seen from the mutated pyrimidine nucleotide, induced by each NA. Only NAs are shown that caused a significant number of SBS (Fig. 1B). The counts were corrected for the background mutagenesis by subtracting the profile of the average control clone from each treatment profiles before counting the mutation types. P-values were calculated by the fisher test and fdr-corrected. * = $p < 0.05$. ** = $p < 0.0001$. **B)** The reference bases of the double base substitutions (DBS) found in clones exposed to the four compounds that induced DBS. 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. **D)** Cosine similarities of all the NA-induced mutational profiles shown in C and the control profile of untreated clones. **E)** Top, the mutational profile of mutations found in the genome of SARS-CoV-2 viruses in humans after they were treated with molnupiravir. Bottom, the mutational profile induced by molnupiravir found in this manuscript in human HSPCs treated in vitro. 3TC = lamivudine, AZT = zidovudine, BVD = brivudine, DDC = zalcitabine, GCV = ganciclovir, MOV = molnupiravir (NHC), PCV = penciclovir.

two six had non-cyclic sugar-like moieties. None of the mutagenic NAs harbored a phosphate group or had a change in the glycosidic bond, but these characteristics were only present in one NA in our test.

Each antiviral NA induces different mutation types

In order to gain further insight into the mechanism of NA mutagenicity, we analyzed the type of base substitutions that were induced by the six mutagenic NAs (**Fig. 2A**). We found that for ganciclovir, penciclovir, and brivudine the mutation spectra, after background correction (see **Methods**), were significantly different from the primarily C>T mutations that were found in untreated control clones. For these drugs, the type of mutations matched their corresponding nucleotide, i.e., G:C > T:A mutations for guanine analogues ganciclovir and penciclovir and T:A > C:G mutations for thymidine analogue brivudine. Molnupiravir induces both C:G > T:A and T:A > C:G mutations. Zidovudine and zalcitabine induce all six types of mutations to approximately the same extend. We also investigated the type of DBS. Except for 3TC, the majority of DBS were TN>NN DBS (**Fig. 2B**).

When also considering the base preceding and following the mutated base of SBS, a spectrum of 96 trinucleotide changes arises. Except for ganciclovir and penciclovir, all spectra were unique for each NA compound (**Fig. 2C, D**). Zalcitabine and zidovudine did not induce mutations in specific contexts. In contrast, ganciclovir and penciclovir almost exclusively induced CpA>ApA mutations. Exposure to brivudine and molnupiravir also resulted in mutations in specific contexts, although with more sequence variation than ganciclovir/penciclovir. A molnupiravir mutational signature was recently reported in SARS-CoV-2 genomes treated with this drug. It consisted of the same two mutation types but had fewer T>C compared to C>T mutations and the mutations occurred in different contexts (cosine similarity 0.32, **Fig. 2E**).

Genomic distribution of NA-induced mutations only differs for ganciclovir

We have previously reported that ganciclovir-induced mutations display strand asymmetries as well as genomic distribution biases[22]. The mutated guanine of the C:G reference base pairs are enriched on the lagging strand of the DNA, the untranscribed strand of genes, in early replicating regions, and in promoters and exons[22]. In addition, ganciclovir-induced mutations display a specific extended sequence context with among others a depletion of adenines at the -2 position. Penciclovir, which is very similar in structure to ganciclovir, showed similar replication and transcription stand biases of the C>A mutations of ganciclovir, but these were not significant, possibly due to the lower number of mutations that penciclovir pound induced (**Fig. 3A-C**). In addition, penciclovir shares none of the extended context with ganciclovir (**Fig. 3D**). ACV is also structurally similar to ganciclovir, but treatment with ACV does not result in more mutations than background, and these mutations show no such biases. Interestingly, most NA-induced mutations are significantly depleted in early replicating regions and enriched in late replication regions, while for ganciclovir the

opposite is observed (**Fig. 3E**). Similarly, all mutagenic NA induced mutations that were depleted in promoters and exons, except for ganciclovir (**Fig. 3F**).

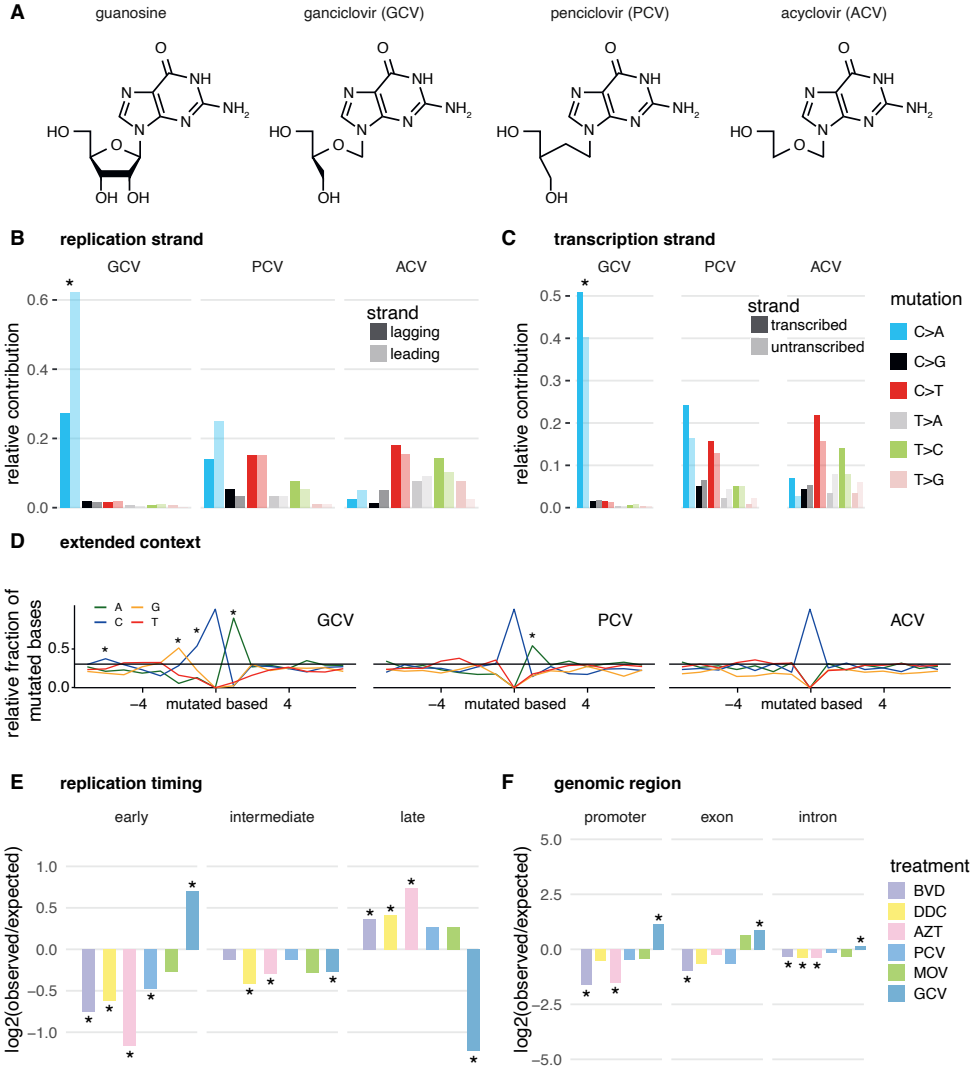


Figure 3. Ganciclovir is the only NA that has strong replication, transcription, and genomic location biases

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 and PCV and ACV are also shown for comparison. **D)** The enrichment or depletion of mutations in regions of early, intermediate, and late replication timing of NAs that induced significant SBS. **E)** The enrichment or depletion of mutations in promoter, exon, or intron 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 one base. ACV is shown for comparison. For this analysis, mutations were split by C>N and T>N mutations per NA. Only the C>N are depicted here.

Discussion

We screened fourteen NAs for mutagenicity in human UCB-derived HSPCs and found evidence for mutagenicity for five out of fourteen tested compounds. Two of these, zalcitabine and zidovudine, miss the 3' hydroxyl group needed for DNA elongation. It seems therefore unlikely that their mutagenicity is induced by incorporation into the DNA and subsequent mismatching. Interestingly, the mutations that these two molecules induced were of all six mutation types and were not enriched in a specific context. Possibly, these drugs induce mutations via other, less direct mechanisms. For example, they might be interfering with polymerases or DNA damage repair proteins that therefore have an altered function. Alternatively, they lead to stress and subsequent stress-induced mutagenesis. A similar general mutagenesis without specific contexts was recently reported in normal human cells after chemotherapy exposure[25].

None of the tested compounds was as mutagenic as the previously reported ganciclovir (GCV)[22], even though we included the structurally very similar penciclovir (PCV) and acyclovir (ACV). ACV was not mutagenic in our screen. Possibly the much lower intracellular half-life of ACV compared to GCV and PCV might play a role in this[26–28]. More importantly, ACV is the only of the three compounds that does not have the 3' hydroxyl group on what normally is the deoxyribose ring of the nucleoside. Therefore, once incorporated, ACV cannot form a bond with another nucleotide and always induces chain termination[16,28]. It can therefore not be mutagenic through incorporation. PCV and GCV have this hydroxyl group and viral polymerases are known to be able to elongate DNA after PCV and GCV incorporation[27,29]. In addition, GCV can also be incorporation into the DNA of a human cancer cell line[23]. Our results suggest that human polymerases can also incorporate PCV without chain termination. But what explains the difference in mutagenicity between PCV and GCV? PCV is phosphorylated by non-infected cells to similar levels as GCV, not explaining the difference[17]. The structural difference between PCV and GCV is only the lack of an oxygen in the acyclic sugar-like moiety, but their biological properties and effectiveness against viruses like varicella zoster virus and cytomegalovirus are very different[27,30]. It is thus likely that this structural difference also explains the difference in the mutagenicity in human uninfected cells. As GCV has a structure that is more similar to guanosine than PCV, it might be incorporated by more human DNA polymerases or might be more efficiently incorporated by a specific DNA polymerase. Future studies should compare the rates of incorporation and chain termination of GCV and PCV by a variety of human DNA polymerases in non-infected cells to confirm our results. Possibly, such a study could also elucidate the reason for the high mutagenicity of GCV, even though it is phosphorylated to a much lower extend (>10x less) in uninfected cells compared to CMV infected cells[17]. Finally, the enrichment of GCV-induced mutations in early replicating regions, and protomers might be explained by such a functional study.

Another NA that was mutagenic in our screen was molnupiravir, which is used to treat severe SARS-CoV-2 infections. Molnupiravir treatment results in an increase of mutations in the SARS-CoV-2 genome in patients[31,32]. It was previously shown to compete strongly with cytosines, but to a lesser extend also with uracils[33]. Once incorporated, either a guanine or an adenine can be incorporated on the opposite strand during replication[33,34]. Indeed, specifically C>T/G>A and to a lesser extend T>C/A>G mutations were found in MOV-treated viral RNA genomes[34]. Interestingly, the mutational profile identified in the genome of SARS-CoV-2 was different to the profile that we report here. Possibly, the limited number of sites that can be mutated in the ~30Kb size of the SARS-CoV-2 genome might result in a different signature than mutations in the ~3Gb human genome, although the authors tried to account for this in the SARS-CoV-2 signature[34]. In addition, the viral mutations were identified in vivo, which means the mutations are under selective pressure and some mutations might be negatively or positively selected, resulting in a different mutational signature. Finally, the functional difference between the SARS-CoV-2 RNA polymerase and the human DNA polymerases might explain this difference. Molnupiravir has the ribose backbone found in RNA nucleosides and is therefore likely more efficiently incorporated into the viral RNA genome than human DNA. In addition, ribonucleotides can be incorporated in the human DNA, but this happens at a low rate, and possibly more often by specific DNA polymerases, e.g., translesion polymerases[35,36]. Further experiment, for example with DNA polymerase knock-out cells, might teach us what human DNA polymerases are involved in the incorporation of molnupiravir.

Our data also show that antiviral NAs do not induce SVs, even though the treatment with GCV can induce yH2AX foci in vitro[22]. In addition an increased level of aneuploidy was previously reported in UCB-derived T cells of HIV-infected pregnant women treated with zidovudine[37]. Possibly, cells which such damage are not able to clonally expand and are therefore missed in our study. In vitro treatment followed by direct DNA amplification from a single cell might prevent this bias. For example, primary template-directed amplification (PTA) could be used for this[38,39].

Our work emphasizes the importance of thorough genotoxicity testing of human drugs. Five out of fourteen tested NAs were mutagenic. Not in all cases does the structure of a nucleoside analogue predict whether it is mutagenic in human cells. For example, NAs without a 3' hydroxyl group or with a 2' hydroxyl group can both be mutagenic, while both characteristics are not found in normal deoxyribonucleosides. Finally, although the number of mutations induced by one dose of some of the drugs is not high, long-term exposure to the treatment may have important repercussions for the exposed healthy cells. General genotoxicity testing using a method based on WGS analysis, such as the one used here, is therefore important to comprehensively assess drug safety with respect to mutagenicity.

Methods

Collection of cord blood samples

Umbilical cord blood samples were collected through the WKZ maternity ward in Utrecht in accordance with the Declaration of Helsinki. 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

The mononuclear cell fraction was isolated from the cord blood sample using LeucosepTM tubes (Greiner Bio-One) and snap-frozen in DMSO. On the day of the experiment, mononuclear cells were thawed in pre-warmed IMDM media supplied with 10% FBS. Cells were washed two times with IMDM+10% FBS at 350g 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 (MACS) with anti-CD34 magnetic beads (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 $\times 10^5$ to 1 $\times 10^5$ cells/mL in a 12-well plate filled with 2 mL medium, respectively. After 24h of recovery, cells were exposed to the nucleoside analog 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 spun down at 350g 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 HSPCs were: Lineage-CD34+CD38-CD45RA-CD90+CD11c-CD16- or Lineage-CD34+CD38-CD45RA-CD49f+CD11c-CD16-. Flow cytometry data were analyzed using the 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 (described above). After 3-6 weeks of culture at 37°C and 5% CO2 5% O2, confluent wells were collected for DNA isolation and WGS.

Whole-Genome Sequencing, read alignment, mutation calling, and filtering

HSPC clones were sequenced with a median genome target coverage 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 rounds 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 variants 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 of 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.
Where 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. Although processing the clones treated with EMT and ACV in pairs instead of three clones, the mutation load was not significantly higher than control clones, indicating that this difference did not impact the outcome of the study.

Structural variant and copy number variant calling

Somatic structural variants (SVs) and copy number variants (CNVs) were called using the gridss-purple-linx pipeline v1.3.2 (https://github.com/hartwigmedical/hmftools) with options '—amber_tumour_only "true" –cobalt _tumour_only "true" –purple_tumour_only "true"', using another clones of the same patient as a reference.

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. SBS had an average variant allele-frequency (VAF) of 0.5 (**Fig. S1A**). 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 (**Fig. S1B**). 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[40].

The MutationalPatterns package was also used for determining cosine similarities between mutational profiles, to calculate replication and transcription strand biases, and genomic and replication timing enrichment/depletion.

The molnupiravir-induced signature in SARS-CoV-2 genome derived from data of Alteri et al.[31] were downloaded from Sanderson et al.[34].

For mutation types (N>N) and mutational profiles (96-trinucleotide profiles N[N>N]N), background correction was done by subtracting the average number of each mutation type/context in the control clones from the average number of the same mutation types/contexts in the treated clones. The was repeated for each treatment separately.

Mitochondrial DNA coverage

The coverage depth of the mitochondrial genome was performed using a previously published pipeline[41], 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 indicated potential technical artifacts (n=1).

Statistics

All p-values were FDR-corrected. The significance of the increase in the number of indels and SBS found in NA-treated HSCP clones compared to control, untreated HSPC clones was done by the Wilcoxon-test. For the number of DBS, the Fisher's exact test was used on the counts of clones with or without any DBS (>0). These values were then FDR corrected. To determine the difference in the mutation types

(N>N), the mean number of C>N and T>N mutations were compared between NA-treated and control clones by fisher's exact test.

The enrichment/depletion of extended contexts was done by fisher's exact test on the number of each of the four nucleotides found at a relative certain position (e.g. 2 base pairs downstream of the mutated base) in the NA-treated and the control clones. This was repeated for each treatment, for T>N and C>N mutations separately, and for the -10 to +10 position relative to the mutated base.
The significance of transcription strand bias and replication strand bias were calculated by MutationalPatterns' "strand_bias_test" and for replication timing enrichment and genomic enrichment MutationalPatterns' "enrichment_depletion_test" was used.

Acknowledgments

This work was funded by an ERC Consolidator grant to R. van Boxtel from the European Research Council (ERC; no. 864499). In addition, all authors were supported by the Oncode institute. The authors want to thank the Hartwig Medical Foundation for facilitating the low-input whole-genome sequencing.

Author contributions

F.P. and J.K. wrote the manuscript. F.P. performed the majority of the experiments with input from R.v.B. N.G, L.T., and L.D. assisted in the experimental work. J.d.K. performed the bioinformatic analysis and prepared the figures. M.R. and F.M. assisted in running pipelines. All authors approved the manuscript for publication.

References

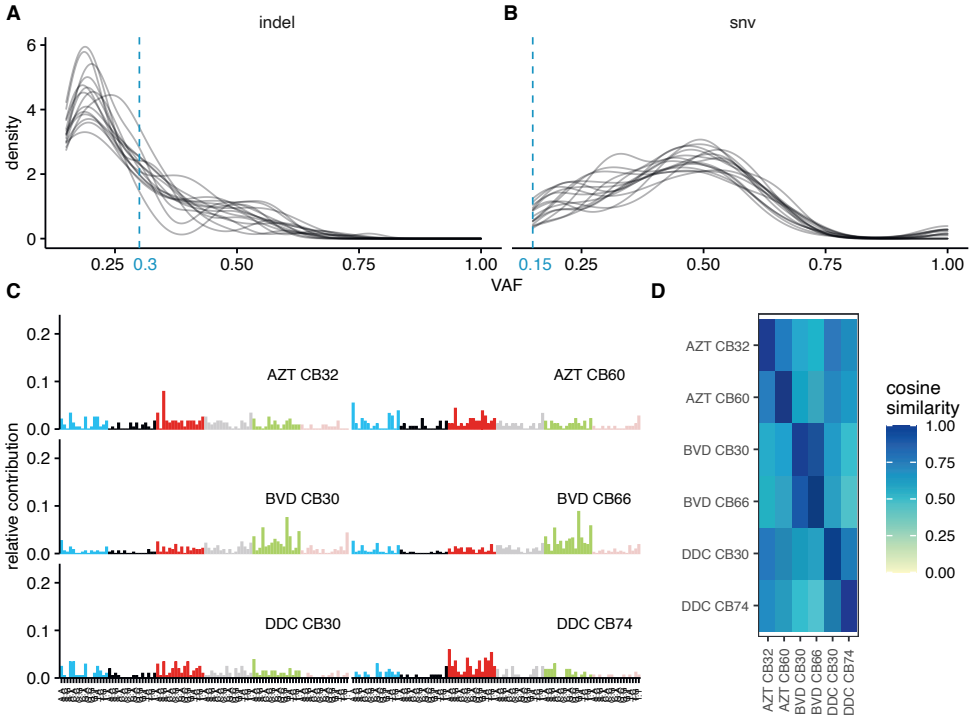
1. P, L. A., M, B. T. & Michael, B. Progress and Challenges in the Prevention, Diagnosis, and Management of Cytomegalovirus Infection in Transplantation. Clin Microbiol Rev 34, 10.1128/cmr.00043-19 (2020).
2. Griffiths, P. & Reeves, M. Pathogenesis of human cytomegalovirus in the immunocompromised host. Nat Rev Microbiol 19, 759–773 (2021).
3. Heaton, P. M. Challenges of Developing Novel Vaccines With Particular Global Health Importance. Frontiers in Immunology vol. 11 Preprint at https://doi.org/10.3389/fimmu.2020.517290 (2020).
4. 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).
5. Andrei, G. & Snoeck, R. Advances and Perspectives in the Management of Varicella-Zoster Virus Infections. Molecules 26, 1132 (2021).
6. Birkmann, A. & Zimmermann, H. HSV antivirals – current and future treatment options. Curr Opin Virol 18, 9–13 (2016).
7. Zou, R. et al. Antiviral Efficacy and Safety of Molnupiravir Against Omicron Variant Infection: A Randomized Controlled Clinical Trial. Front Pharmacol 13, (2022).
8. Fischer, W. A. et al. A phase 2a clinical trial of molnupiravir in patients with COVID-19 shows accelerated SARS-CoV-2 RNA clearance and elimination of infectious virus. Sci Transl Med 14, eabl7430 (2023).
9. Pruijssers, A. J. & Denison, M. R. Nucleoside analogues for the treatment of coronavirus infections. Curr Opin Virol 35, 57–62 (2019).
10. Delma, F. Z. et al. Molecular mechanisms of 5-fluorocytosine resistance in yeasts and filamentous fungi. Journal of Fungi vol. 7 Preprint at https://doi.org/10.3390/jof7110909 (2021).
11. Galmarini, C. M., Mackey, J. R. & Dumontet, C. Nucleoside analogues and nucleobases in cancer treatment. Lancet Oncol 3, 415–424 (2002).
12. 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).
13. Kausar, S. et al. A review: Mechanism of action of antiviral drugs. International Journal of Immunopathology and Pharmacology vol. 35 Preprint at https://doi.org/10.1177/20587384211002621 (2021).
14. A, J. S., Eisuke, M., William, D., Phillip, F. & Jianming, H. Noncompetitive Inhibition of Hepatitis B Virus Reverse Transcriptase Protein Priming and DNA Synthesis by the Nucleoside Analog Clevudine. Antimicrob Agents Chemother 57, 4181–4189 (2013).
15. 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).
16. Talarico, C. L. et al. Acyclovir Is Phosphorylated by the Human Cytomegalovirus UL97 Protein. Antimicrob Agents Chemother 43, 1941–1946 (1999).
17. Zimmermann, A. et al. Phosphorylation of aciclovir, ganciclovir, penciclovir and S2242 by the cytomegalovirus UL97 protein: a quantitative analysis using recombinant vaccinia viruses. Antiviral Res 36, 35–42 (1997).
18. Furman, P. A. et al. Phosphorylation of 3'-azido-3'-deoxythymidine and selective interaction of the 5'-triphosphate with human immunodeficiency virus reverse transcriptase. Proceedings of the National Academy of Sciences 83, 8333–8337 (1986).
19. 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).
20. Brambilla, G. & Martelli, A. Update on genotoxicity and carcinogenicity testing of 472 marketed pharmaceuticals. Mutation Research/Reviews in Mutation Research 681, 209–229 (2009).
21. Barrow, P. & Clemann, N. Review of embryo-fetal developmental toxicity studies performed for pharmaceuticals approved by FDA in 2018 and 2019. Reproductive Toxicology 99, 144–151 (2021).
22. 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).
23. Fang, H. et al. Ganciclovir-induced mutations are present in a diverse spectrum of post-transplant malignancies. Genome Med 14, 124 (2022).
24. Rosendahl Huber, A. et al. Whole-genome sequencing and mutational analysis of human cord-blood derived stem and progenitor cells. STAR Protoc 3, (2022).
25. 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.
26. Hodge, R. A. & Perkins, R. M. Mode of action of 9-(4-hydroxy-3-hydroxymethylbut-1-yl)guanine (BRL 39123) against herpes simplex virus in MRC-5 cells. Antimicrob Agents Chemother 33, 223–229 (1989).
27. Hodge, R. A. V. & Cheng, Y.-C. The Mode of Action of Penciclovir. Antivir Chem Chemother 4, 13–24 (1993).
28. Elion, G. B. Acyclovir: Discovery, mechanism of action, and selectivity. J Med Virol 41, 2–6 (1993).
29. Chen, H., Beardsley, G. P. & Coen, D. M. Mechanism of ganciclovir-induced chain termination revealed by resistant viral polymerase mutants with reduced exonuclease activity. Proceedings of the National Academy of Sciences 111, 17462–17467 (2014).
30. Hannah, J. et al. Carba-acyclonucleoside antiherpetic agents. J Heterocycl Chem 26, 1261–1271 (1989).
31. Alteri, C. et al. A proof-of-concept study on the genomic evolution of Sars-Cov-2 in molnupiravir-treated, paxlovid-treated and drug-naïve patients. Commun Biol 5, 1376 (2022).
32. Donovan-Banfield, I. et al. Characterisation of SARS-CoV-2 genomic variation in response to molnupiravir treatment in the AGILE Phase IIa clinical trial. Nat Commun 13, 7284 (2022).
33. Malone, B. & Campbell, E. A. Molnupiravir: coding for catastrophe. Nat Struct Mol Biol 28, 706–708 (2021).
34. Sanderson, T. et al. A molnupiravir-associated mutational signature in global SARS-CoV-2 genomes. Nature 623, 594–600 (2023).
35. Koh, K. D., Balachander, S., Hesselberth, J. R. & Storici, F. Ribose-seq: global mapping of ribonucleotides embedded in genomic DNA. Nat Methods 12, 251–257 (2015).
36. Vaisman, A. & Woodgate, R. Ribonucleotide discrimination by translesion synthesis DNA polymerases. Crit Rev Biochem Mol Biol 53, 382–402 (2018).
37. 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).
38. Middelkamp, S. et al. Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox. Cell Genomics 3, 100389 (2023).
39. Gonzalez-Pena, V. et al. Accurate genomic variant detection in single cells with primary template-directed amplification. Proc Natl Acad Sci U S A 118, 1–12 (2021).
40. Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. (2016).
41. Manders, F., van Dinter, J. & van Boxtel, R. Mutation accumulation in mtDNA of cancers resembles mutagenesis in normal stem cells. iScience 25, (2022).

Supplementary Material

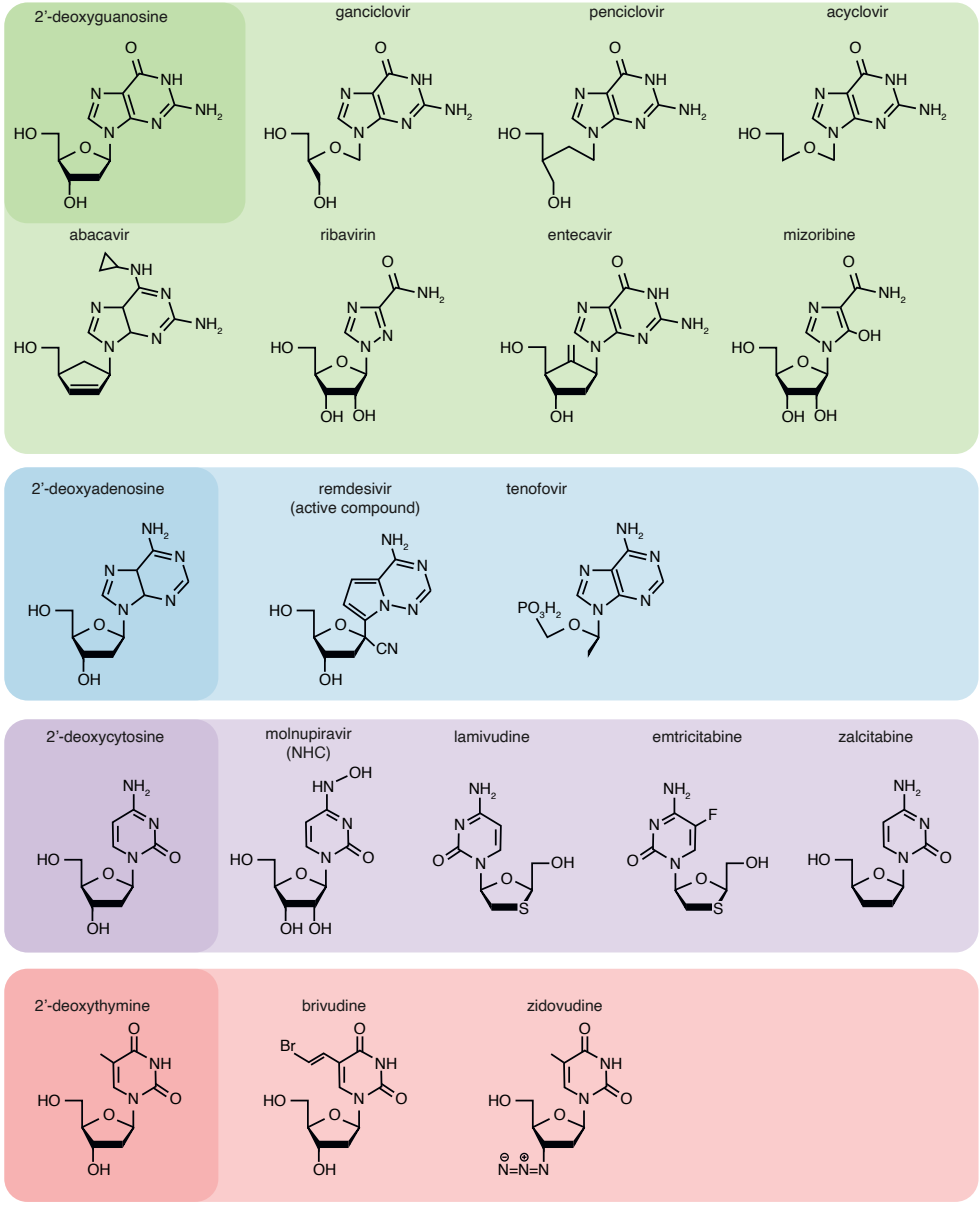
Supplementary Table 1. Technical specification of the tested NAs.

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

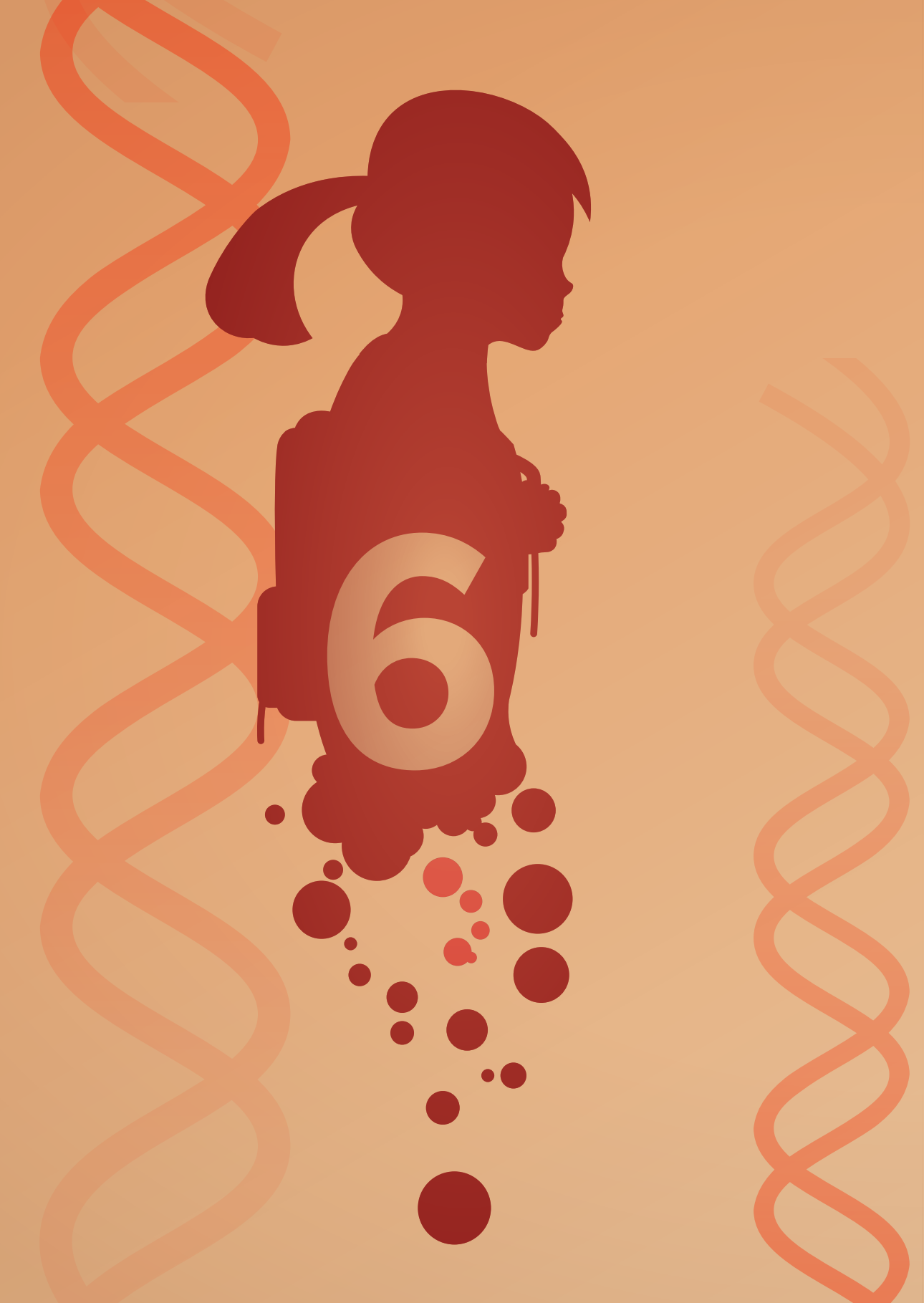


Supplementary Figure 1. Variant allele frequencies cut-offs and profiles of biological duplicates.

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 indels (0.3) is indicated. **B)** similar to A, but for single base substitutions. The VAF cut-off of 0.15 is indicated. **C)** The 96 SBS mutation profiles of zidovudine (AZT), brivudine (BVD) and zalcitabine (DDC) in two independent UCB donors. The profiles of all three clones from each donor treated with one NA were averaged. **D)** The cosine similarity between the profiles showed in C.



Supplementary Figure 2. The structure of the nucleoside analogs tested in this manuscript along with the structure of the corresponding DNA nucleosides.


6

CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing

Jurrian K. de Kanter[1,*], Philip Lijnzaad[1,*], Tito Candelli[1], Thanasis Margaritis[1], Frank C.P. Holstege[1]

[1] Princess Máxima Center for Pediatric Oncology, Utrecht, 3584 CS, The Netherlands
* These authors contributed equally

Nucleic Acids Research, June 21;47(16): e95
doi: 10.1093/nar/gkz543 (2019)

Abstract

Cell type identification is essential for single-cell RNA sequencing (scRNA-seq) studies, currently transforming the life sciences. CHETAH (CHaracterization of cEll Types Aided by Hierarchical classification) is an accurate cell type identification algorithm that is rapid and selective, including the possibility of intermediate or unassigned categories. Evidence for assignment is based on a classification tree of previously available scRNA-seq reference data and includes a confidence score based on the variance in gene expression per cell type. For cell types represented in the reference data, CHETAH's accuracy is as good as existing methods. Its specificity is superior when cells of an unknown type are encountered, such as malignant cells in tumor samples which it pinpoints as intermediate or unassigned. Although designed for tumor samples in particular, the use of unassigned and intermediate types is also valuable in other exploratory studies. This is exemplified in pancreas datasets where CHETAH highlights cell populations not well represented in the reference dataset, including cells with profiles that lie on a continuum between that of acinar and ductal cell types. Having the possibility of unassigned and intermediate cell types is pivotal for preventing misclassification and can yield important biological information for previously unexplored tissues.

Introduction

Single-cell RNA-sequencing (scRNA-seq) is transforming our ability to study heterogeneous cell populations[1–6]. While tools to help interpret scRNA-seq data are developing rapidly[7–15], challenges in data analysis remain[16], with cell type identification a prominent example. Accurate cell type identification is a prerequisite for any study of heterogeneous cell populations, both when the focus is on subsets of a particular cell type of interest or when investigating the population structure as a whole[17–21]. The introduction of single cell RNA sequencing has paved the way for rapidly discovering previously uncharacterized cell types[22–24] and this application too would greatly benefit from efficient identification of known cell types prior to focusing on new types.

Research into tumor composition presents an even more challenging setting, as the RNA expression profile of malignant cells is often different from any known cell type, as well as unique to the patient or even to the biopsy[25,26]. Malignant cells can sometimes be identified in scRNA-seq data[27] but this is not always feasible or even possible, for instance with tumors that do not harbour easily identified copy number variations. In both cases, a first sign of the malignancy of cells in the sample is their imperviousness to classification, simply because their expression profiles do not resemble that of any known, healthy cell type.

Cell type identification in scRNA-seq studies is currently often done manually, starting by identifying transcriptionally similar cells using clustering. This is frequently followed by differential expression analysis of the resulting cell clusters combined

with visual marker gene inspection[4,25,26,28–30]. Such manual cell type identification is time-consuming and often subjective due to the choice of clustering method and parameters for example, or to the lack of consensus regarding which marker gene to use for each cell type. Such analyses are becoming more complex given the fast-expanding catalogue of defined cell types[16]. Canonical cell surface markers are also not always suitable in scRNA-seq studies because the transcripts of these genes may not be measurable in the corresponding cell type owing to low expression or to degradation of the mRNA. This is aggravated by technical difficulties (drop-out) and, more generally, by the poor correlation between protein expression and mRNA abundances[23].

Recently, a number of cell type identification algorithms have emerged to address these problems. Automated methods such as scmap[31] and SingleR[32] base their cell type call on comparisons with annotated reference data using automatically chosen genes that optimally discriminate between cell types. A good cell type identification method should be both sensitive and selective. That is, it should correctly identify as many cells as possible, while not classifying cells when based on insufficient evidence. If the cell being identified is of a type that is not represented in the reference, such misclassification can easily occur. This is a concern when studying malignant cells which are often too heterogeneous to include in the reference data. To avoid overclassification, methods such as scmap[31] therefore leave cells unclassified if they are too dissimilar to any reference data.

Both the complete lack of classification as well as overclassification is unsatisfactory. For example, if the evidence for a very specific cell type assignment such as effector CD8 T-cell is not strong enough, a more general, less specific assignment such as T-cell may still legitimately be made and might still be useful. The reason for such an intermediate cell type assignment could be that the correct T-cell subtype is not part of the reference dataset, or even that there is insufficient read-depth for the more specific call to be made. An even more interesting case is that of cells that are biologically of an intermediate type such as differentiating cells or cells undergoing transdifferentiation.

Here we present CHETAH (CHaracterization of cEll Types Aided by Hierarchical classification), an algorithm that explicitly allows the assignment of cells to an intermediate or unassigned type. The unassigned and intermediate types are inferred using a tree that is constructed from the reference data and which guides the classification. CHETAH's classification is a stepwise process that traverses the tree and, depending on the available evidence, ends at one of the reference cell types or halts at the unassigned or one of the intermediate types. CHETAH is able to correctly classify published datasets and, in comparison to other methods, performs equally or better when considering cells whose type is represented in the reference data. For cells of an unknown type, CHETAH is more selective, yielding a classification that is as fine-grained as is justified by the available data. The benefit of calling unassigned and intermediate types is highlighted in several tumor datasets, showing CHETAH is consistently selective. This makes CHETAH a powerful tool for identifying cells

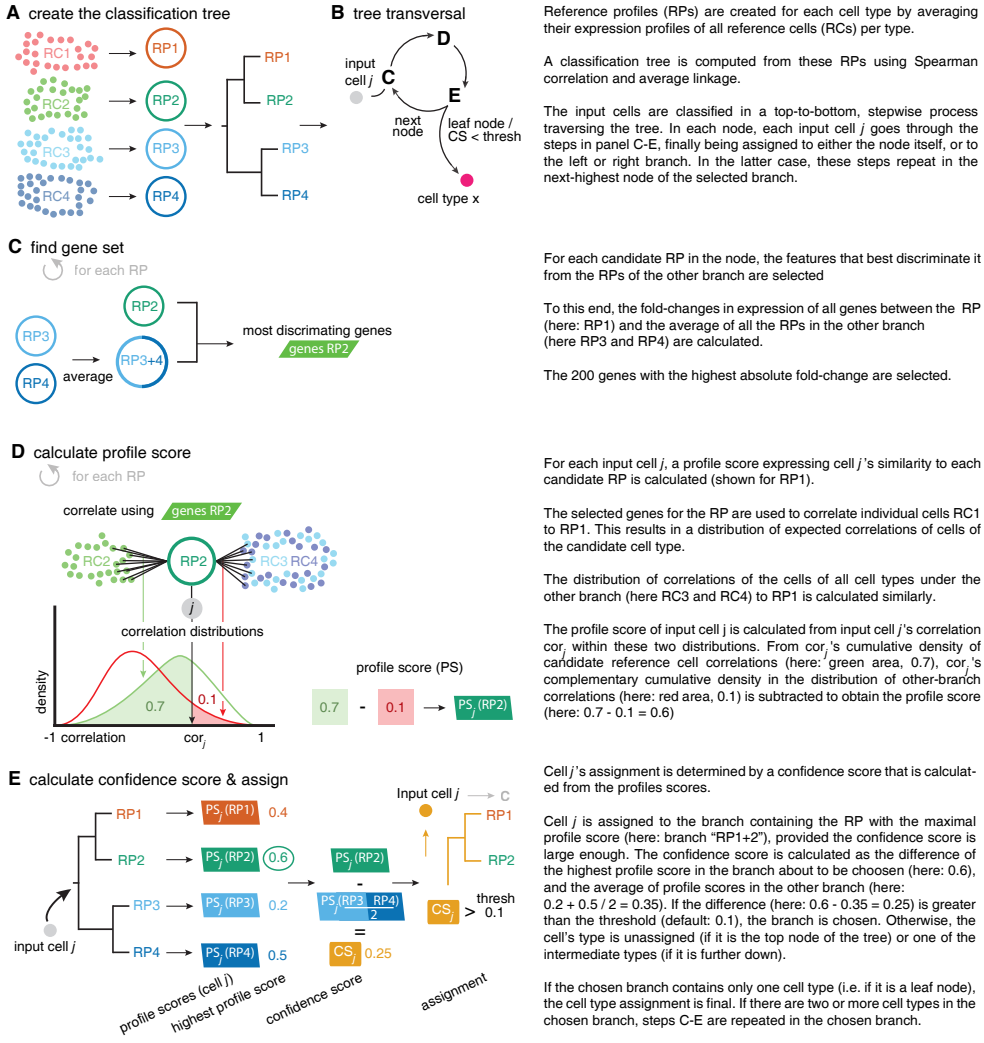
that are not in the reference, such as malignant tumor cells, novel or intermediate cell types. The latter is shown in an analysis of published pancreas datasets, where a manifest expression gradient of cells with types varying between acinar and duct cell is described. CHETAH is implemented in R[33], is available at github.com/jdekanter/CHETAH, and has been incorporated in Bioconductor[34] release 3.9. It comes with an extensive Shiny[35] application that makes exploration of the cell type identification process and the gene expression differences that support the classification very intuitive. CHETAH has been created bearing tumor analyses in mind, but as is demonstrated, it also complements existing methods for exploring previously uncharacterized non-cancerous tissues and cell types.

Methods

An outline of the CHETAH algorithm is depicted in **Fig. 1**. The method requires a reference scRNA-seq dataset, annotated for cell type. Throughout this study, the reference dataset is completely independent from the dataset that is being classified. First a hierarchical classification tree is constructed from the reference scRNA-seq data (**Fig. 1A**). Each input cell is classified individually by traversing the tree (**Fig. 1B**). At each step of the process the cell to be classified is correlated to the expression profiles of the reference data cell types. This is done by first selecting the set of genes that best discriminates each reference cell type from all the cell types, collectively, in the opposite branch of the tree (**Fig. 1C**). The input cell correlation to a reference cell type is compared to the distribution of correlations of the reference cells to assess whether there is enough evidence to allow this cell to take the next step (**Fig. 1D-E**). If the confidence threshold is not passed, further classification of the cell stops and it is marked as unassigned if the evidence runs out at the top of the tree, or as intermediate if this happens within the classification tree. Classification also stops when a cell reaches one of the leaves of the tree, yielding assignment to a specific cell type.

Reference data

In order to classify input cells CHETAH requires scRNA-seq reference data along with cell type labels for each reference cell. Here, the reference dataset is always a completely independently generated dataset, from a different study and in several cases using a different scRNA-seq platform. The reference data needs to be normalised to an identical total number of transcripts per cell and should be expressed in log-scale. Malignant cells are best left out of the reference because they are too ill-defined and too patient-specific[36]. In all the reference datasets used here, such cells cluster by patient whereas non-malignant cells largely cluster by cell type. The reference must contain at least 10 cells per cell type to adequately represent its transcriptional program as well as its variance (**Fig. 3C**). More than 10 cells per reference cell type improves performance. More than 200 cells per cell type is superfluous. Since this also increases the computational burden it is useful to restrict the number of cells per reference cell type to a maximum of 500. This does not restrict the number



Reference profiles (RPs) are created for each cell type by averaging their expression profiles of all reference cells (RCs) per type.

A classification tree is computed from these RPs using Spearman correlation and average linkage.

The input cells are classified in a top-to-bottom, stepwise process traversing the tree. In each node, each input cell *j* goes through the steps in panel C-E, finally being assigned to either the node itself, or to the left or right branch. In the latter case, these steps repeat in the next-highest node of the selected branch.

For each candidate RP in the node, the features that best discriminate it from the RPs of the other branch are selected

To this end, the fold-changes in expression of all genes between the RP (here: RP1) and the average of all the RPs in the other branch (here RP3 and RP4) are calculated.

The 200 genes with the highest absolute fold-change are selected.

For each input cell *j*, a profile score expressing cell *j*'s similarity to each candidate RP is calculated (shown for RP1).

The selected genes for the RP are used to correlate individual cells RC1 to RP1. This results in a distribution of expected correlations of cells of the candidate cell type.

The distribution of correlations of the cells of all cell types under the other branch (here RC3 and RC4) to RP1 is calculated similarly.

The profile score of input cell j is calculated from input cell *j*'s correlation cor_j within these two distributions. From cor_j's cumulative density of candidate reference cell correlations (here: green area, 0.7), cor_j's complementary cumulative density in the distribution of other-branch correlations (here: red area, 0.1) is subtracted to obtain the profile score (here: 0.7 - 0.1 = 0.6)

Cell *j*'s assignment is determined by a confidence score that is calculated from the profiles scores.

Cell *j* is assigned to the branch containing the RP with the maximal profile score (here: branch "RP1+2"), provided the confidence score is large enough. The confidence score is calculated as the difference of the highest profile score in the branch about to be choosen (here: 0.6), and the average of profile scores in the other branch (here: 0.2 + 0.5 / 2 = 0.35). If the difference (here: 0.6 - 0.35 = 0.25) is greater than the threshold (default: 0.1), the branch is chosen. Otherwise, the cell's type is unassigned (if it is the top node of the tree) or one of the intermediate types (if it is further down).

If the chosen branch contains only one cell type (i.e. if it is a leaf node), the cell type assignment is final. If there are two or more cell types in the chosen branch, steps C-E are repeated in the chosen branch.

Figure 1. The CHETAH algorithm.

of cells that can be analysed in the query dataset. The selection of genes from the reference dataset for classification at each step (**Fig. 1B**) is aimed at finding those with the highest discriminatory power. When using a reference dataset with a high drop-out rate, i.e. low transcript coverage per cell, it is advocated to remove highly expressed genes such as ribosomal protein genes from reference datasets beforehand, since their drop-out can reduce classification accuracy. For the sake of uniformity, ribosomal protein genes were removed here from all reference datasets although this only increased classification accuracy in one case (**Fig. 3C**).

Unless stated otherwise, the reference dataset used here, called 'Tumor ref', consists of a combination of datasets of Colorectal[37], Breast[38], Melanoma[25] and Head-Neck tumors[26]. The data for all these studies was generated using the Smart-seq2 platform.

The cell types of the reference dataset were based on published manual classification of cell clusters using marker gene expression. The Melanoma and Head-Neck studies discuss the T-cells in terms of their *CD4+*, *CD8+* and *T-reg* subtypes but not all of these labels are available for all the cells in the online material of these publications. These reference cells were therefore classified manually using the same marker genes as used in the publications. Cells of a dataset being classified are of course excluded from the reference. In other words, all results reported are by classification using reference datasets completely independent from the query dataset. When comparing CHETAH and SingleR[32] results, the latter was run with averaged single-cell data because SingleR uses bulk, rather than single-cell expression profiles as its reference. Further details of the datasets used in this work, including pre-processing, are given in **Table 1**.

Table 1. Datasets used in this study.

name (publication)	protocol	nr. of healthy cells	nr. of tumor cells	nr. of cell types	pre-processing after download	biopsies	nr. of donors
Melanoma (25)	Smart-seq2	3262	1251	9	discarded cells without annotation	melanoma	19
Head-Neck (26)	Smart-seq2	3345	2215	12	discarded cells without annotation	primary head and neck squamous cell carcinoma	18
Colorectal (37)	Smart-seq2	272	92	7	no	colorectal cancer	11
Breast (38)	Smart-seq2	198	317	3	no	breast cancer	11
Tumor ref. (25,26,37,38)	Smart-seq2	6122	none	12	combined Melanoma, Head-Neck , Breast and Colorectal datasets, discarding malignant cells.	Detailed above	19 18 11 11
Ovarian (19)	InDrops	2814	300	9	no	ovarian cancer ascites	4
PBMC (28)	10X Genomics	68579	none	16	no	healthy PBMC cells	1
CBMC (41)	Drop-seq	7830	none	13	as described (41)	cord blood CBMCs	unknown
Pancreas1 (17)	inDrops	8569	none	14	no	healthy pancreas	4
Pancreas2 (43)	CEL-seq2	2292	none	9	no	healthy pancreas	4

The PBMC and CBMC datasets were labelled identically to ensure comparability of annotated cell types

Classification tree

The first step is to create a reference profile (RP) for each cell type in the set of reference cells by averaging, for each cell type, the logged gene expression over all cells of that type (**Fig. 1A**). The RPs are subsequently clustered hierarchically using

Spearman correlation and average linkage to obtain the classification tree.

Hierarchical classification

The classification of input cells proceeds in a stepwise fashion, from the root to the leaves of the classification tree. At each step, the branch is selected that contains the reference cell type most likely to be the correct one, but the classification stops if the confidence in this decision becomes too low (see confidence score below). As described under profile score, the choice of the most likely cell type and therefore which branch to choose, is based on the cell's similarities to each of the individual RPs under each branch. The similarity of a cell to a RP under consideration (called the *candidate* RP), in the branch under consideration (called the *candidate branch*), is always in relation to all the RPs under the (so-called) *other branch*. During the classification process, only the leaf node data (i.e. from all cells of a particular reference cell type) are used. Any details of the tree topology under either branch are ignored, i.e. no hypothetical expression profiles are inferred for the intermediate tree nodes. After calculating the cell's similarities to all RPs under both branches, the cell is assigned to the branch that contains the cell type to which it is most similar, provided the evidence is strong enough based on the confidence score.

Feature selection

The similarity of a cell to a reference profile is based on their Spearman correlation. This choice is based on its identical performance to other correlation methods (**Fig. S1A**) and on the fact that there is no assumption about the underlying distribution. The correlation is calculated using the subset of genes that best discriminates between the *candidate* RP in the *candidate branch*, and the average expression profile of the *other branch* as a whole. (The latter is calculated as the mean of all RPs under that branch). The selection of the best subset of genes, a process known as feature selection, is not critical and good results are achieved when simply using the 200 genes that have the largest absolute fold-change between the *candidate* RP and the average expression profile of the *other branch*. This choice is based on a variety of parameter sweeps and shown in **Fig. S1**. It is important to note that the feature set, i.e. the subset of genes used to calculate similarities, is different for each RP and for each node of the classification tree. Many different feature selection methods work well (**Fig. S1**). The use of different discriminatory gene sets at each decision node and for each RP is an important, novel aspect of the method.

Similarities

The similarity of a cell to a RP in the *candidate branch* is of course reflected in their correlation, but the values of these correlations to the various RPs cannot be directly used for comparisons. The reason is that the subset of genes used for each correlation is generally different for each RP and for each node. The similarity of an input cell j to *candidate* RP x is therefore cast in relative terms by using the *cumulative probabilities* of this correlation within two different distributions of correlations. The first one is

the distribution of self-correlations, that is, the distribution of the correlations of the individual cells constituting the *candidate* RP to that *candidate* RP itself. These self-correlations represent the typical correlation values for a cell that is really of that type. The second distribution is that of the non-self correlations. They are the correlations, again to the *candidate* RP, of all the individual reference cells under the *other branch*. They represent the correlation values that can be expected for cells that are not of any type under the *candidate branch*. By contrasting the two cumulative probabilities a profile score is obtained that robustly points the way through the classification tree.

Profile scores

The two *cumulative probabilities* just defined are used to define the profile score $P_x(j)$, representing cell *j*'s similarity to *candidate* RP *x*, as follows:

$$P_x(j) = F_c\,(r_s(j,x)\,) - [\,1 - F_o\,(r_s(j,x)\,)\,] \qquad [1]$$

with

$r_s(j,x)$	the Spearman correlation of input cell *j*'s expression with candidate reference profile *x*
$F_c\,(r_s(j,x))$	the cumulative probability of *j*'s correlation within the distribution of self-correlations $r_s(k,x)$, that is, of all reference cells *k* of type *x* with their 'own' candidate reference profile *x*
$F_o\,(r_s(j,x)\,)$	the cumulative probability of *j*'s correlation within the distribution of correlations $r_s(l,x)$ of all reference cells *l* under the *other branch*, again with reference profile *x*

The profile score $P_x(j)$ has a value between 1 and -1 and is, in a particular node, a measure for the likelihood that cell *j* is of type *x*. A value of 1 means that cell *j* is much more likely to be of type *x* (and therefore belong to its branch) rather than any of the types in the *other branch* and, conversely, -1 represents the lowest likelihood of this being so, and therefore cell *j* is much more likely to belong in the *other branch*. In each node, one set of genes is selected for each RP under that node. This gene set is used for all the correlations (of both input and reference cells) needed to calculate the profile scores. Note that due to the different gene subsets used in each step of the tree traversal, the most similar RP for a cell may change during the steps of the classification process. E.g., during the first few steps a cell that in reality is of type *CD4* T-cell could initially, and incorrectly, appear more similar to a *CD8* T-cell than to the expected *CD4* T-cell type. This would however still lead to the correct branch choice, namely that of all T-cells. In later steps the similarity to the actual *CD4* T-cell type would become strongest, guiding the cell to a correct final *CD4* T-cell label.

Confidence score

Each input cell is assigned to the branch containing the candidate reference cell type for which it has the highest profile score. This assignment represents the choice between the left and right branch, but a key design goal of the algorithm is its ability to stop classification at an intermediate node. The choice for each cell *j*, between stopping classification or continuing to the next round, is based on its confidence score *C(j)* defined as

$$C(j) = P_{max(j)} - mean(P_{o(j)}) \quad [2]$$

with $P_{max}(j)$ the highest profile score for cell j in the branch about to be chosen and $mean(P_o(j))$ the mean of the profile scores in the *other branch*, i.e. the branch not containing the reference profile having the highest profile score (**Fig. 1C**). Expression [2] is always positive because branches leading to a negative score are by definition never chosen by the algorithm. The confidence score is a measure for the evidence to assign a cell to a branch, with 2 representing maximal evidence, and 0 representing no evidence. The confidence score has an easy explanation. If it is close to 0, the best candidate cell type in the branch about to be chosen is as good as the average of the cell types in the *other branch*. This implies that there is no basis to justify the choice between either branch, so none should be taken and classification of the cell should therefore stop in the current node. In contrast, a large score represents good support to continue the classification because there is a cell type in the *candidate branch* that has a much better profile score than the average profile score of the *other branch*. By default, cells are assigned to the branch if the confidence score is higher than 0.1, but different values can also be specified in the algorithm's parameters. Cells that remain in a non-leaf node of the tree are called unassigned or of intermediate type whereas cells assigned to a leaf-node are of a final type. The labels for the intermediate types are generated automatically (Node1, Node2, etc.) but biologically meaningful names such as T cell can often readily be given. By choosing a cut-off greater than 0.1, only the more confident calls will be made, hence more cells will be labelled as being unassigned or of intermediate type. Conversely, by lowering the confidence cut-off, the algorithm will classify more cells to a final type, however such calls are supported by less evidence. A cut-off of 0.0 forces the method to classify all cells to a final type, as is exemplified later. The above stepwise calculations of the profile scores and confidence scores yield an elegant and, importantly, transparent algorithm.

Parameters

CHETAH comes with an extensive Shiny[35] application, is implemented in R[33], is available at github.com/jdekanter/CHETAH and has been incorporated in Bioconductor[34] release 3.9. Easily selectable parameters include the choice of correlation measure (default: Spearman), the discriminatory gene set selection method (default: the 200 genes with largest absolute expression difference between reference cell type under consideration and all reference cell types in the opposite

branch), the hierarchical clustering method (default: Spearman, average linkage) and the confidence score threshold (default: 0.1). These default settings are uniformly applied throughout this study.

Results

The CHETAH algorithm is summarized in the first paragraph of the Methods section and in **Fig. 1**. The method makes use of a reference dataset with cell type annotations. Throughout this study the reference dataset is always a completely independently generated dataset, from a different study and in several cases using a different scRNA-seq platform. Reference cell types are hierarchically clustered into a classification tree which guides the cell type identification process (**Fig. 1A**). The classification tree aids cell type identification but is not intended as a recapitulation of cell taxonomy. The cells to be classified are shunted from the root of this tree to its leaves (**Fig. 1B**), but only to the most specific tree node that is still supported by the available evidence, as quantified by a robust measure of confidence (**Fig. 1D-E**). Confidence is based on passing a threshold. This is determined by the degree to which the input cell's correlation to a reference cell type fits with the distribution of correlations of reference cells of the same type and contrasted with the degree to which it fits with that of other reference cells to this type (**Fig. 1D-E**). The genes on which the classification is based are selected to be those that are most discriminatory for each step in the classification (**Fig. 1C**). This too is an important aspect of the method. Many parameters such as correlation measure, number of discriminatory genes, number of cells per reference type, etc., are selectable by the user (Methods) and the choice for the single set of default parameters used throughout this study is explained in the Methods. Cells for which classification confidence runs out are typically of a type that is not present in the reference dataset, and are said to be either unassigned or of an intermediate type. Intermediate entails that classification has halted at a node due to lack of confidence to proceed. Unassigned entails that this already occurred at the first step in the classification tree. Note that there are several intermediate types, each corresponding to one of the internal nodes of the classification tree.

CHETAH's accuracy is investigated by comparing its classifications with published cell type labels. The aim is to reproduce these using only the reference data. The reference datasets used here are always from a source that is completely independent of the query dataset, ensuring that the reported accuracies do not reflect bias from over-fitting. Since the accuracy might be lower if the scRNA-seq technology of the input data and the reference differs, cross-platform results are also examined. CHETAH is subsequently compared to other cell type identification methods and the effectiveness of the intermediate cell type assignments is also demonstrated in an analysis of previously published pancreas datasets. For an overview of the datasets see **Table 1**.

Accuracy

The performance of CHETAH is first evaluated by applying it to Melanoma[25] and Head-Neck[26] cancer datasets. The classifications of the these datasets is shown in **Fig. 2** and **S2**, summarized in **Table 2**. Throughout this study all classification results are obtained by applying CHETAH on a new query dataset, with a completely different, independent dataset as reference. The reported results are therefore without bias towards the query dataset, as could be the case if reference and query datasets are the same. Since the reference datasets do not contain malignant tumor cells, such cells should not be classified to any final type, but as unassigned or any of the intermediate types instead. CHETAH correctly classifies practically all (mean > 99%) malignant cells as unassigned or intermediate types. Note that in the published data the classifications were manual while the identification of tumor cells was based on estimated copy number variations. In contrast, CHETAH's type assignments are fully automatic and the aberrant nature of the malignant cells is indicated by their classification as unassigned or intermediate. This selectivity is an important quality of the method, essential for preventing the type of misclassification that readily occurs when methods forcefully assign every cell to a type regardless of the evidence. Selectivity is especially relevant when dealing with tumor samples, as well as with samples containing cell types not present in the reference dataset.

Table 2. Percentages of cell type labels as inferred by CHETAH, compared with the published cell types.

Reference dataset	input dataset	% correctly unassigned	% identical final type	% intermediate type, same lineage	% intermediate type, different lineage	% Incorrectly unassigned	% different final type
Tumor ref. without Melanoma	Melanoma (25)	99.5	86.9	4.1	0.8	0.5	7.7
Tumor ref. without Head-Neck	Head-Neck (26)	89.2	71.2	14.6	1.1	1.9	11.2
Tumor ref.	Ovarian (19)	99.3	79.9	9.0	0.7	6.8	3.6

The reference and input datasets are shown. For an overview of the datasets see Table 1. When classifying the Melanoma and Head-Neck datasets, these datasets are left out of the 'Tumor ref.' reference, as indicated. The column correctly unassigned shows the percentage of cells of a type that was absent from the reference that were classified as unassigned or any of the intermediate types. The other columns refer to sample cells of a type represented in the reference that should therefore be assigned and contain percentages of cells of final or intermediate type, summing to 100%. The term lineage refers to the classification tree determined by CHETAH.

CHETAH classifies the majority (mean 79%) of non-malignant cells the same as in the original publication. Of the cells classified differently, the majority (mean 61%) are classified as an intermediate type. In the inferred classification tree these intermediate assignments are overwhelmingly in the correct classification lineage (85%, 91% and 95% for Melanoma, Head-Neck and Ovarian respectively). Only a

small number of cells are labelled differently by CHETAH. For many of them there is in fact strong evidence from established marker gene expression that the assignment from CHETAH is correct (**Fig. S4**, **S5**). Taken together, these results show that the selective approach works well. Cells of an established type that are present in the reference dataset are classified correctly. Samples cells of a new or aberrant type, not represented in the reference dataset are either not assigned to a type or are classified as an intermediate type, an outcome that should indeed be regarded as a pointer for a more detailed inspection.

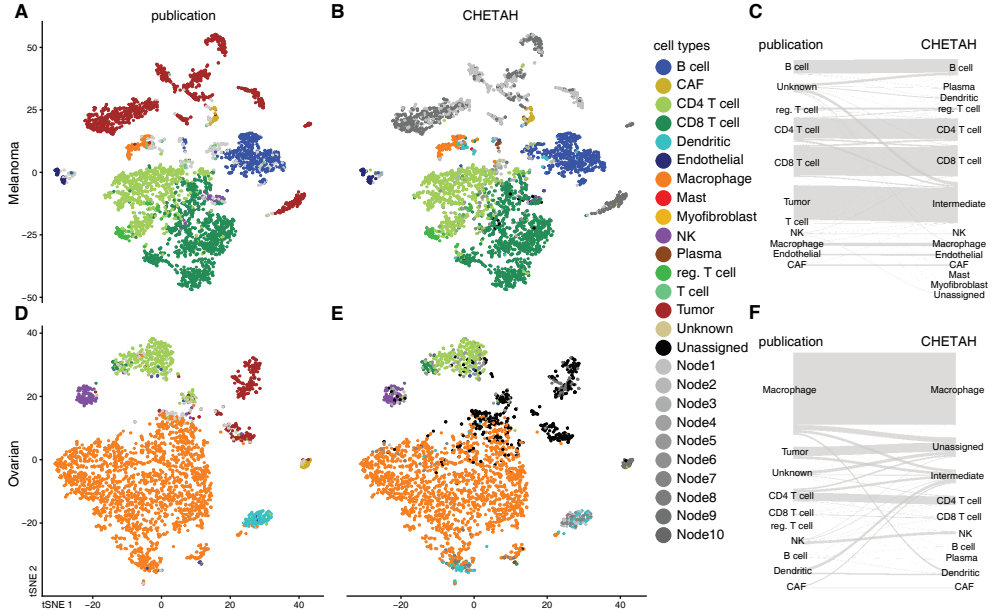


Figure 2. CHETAH's classification of two tumor sample datasets is nearly identical to the published manual classification. The t-SNE plots depict each cell as a dot, with the colours representing the inferred cell type shown in the legend. Gray colours indicate intermediate cell types which are labelled automatically as Node1, Node2, etc. For the corresponding classification trees see Figure S2. Rows of panels: datasets classified (Melanoma: Tirosh et al.[25]; Ovarian: Schelker et al.[19]); columns: classification method. For an overview of the datasets see Table 1.

Cross-platform classification

The data from the Melanoma and Head-Neck studies were obtained using Smart-seq2[39] and were also classified using reference data originating from the same platform. To evaluate CHETAH's performance across platforms, an Ovarian dataset[19] produced on the inDrops platform[40] was analysed with CHETAH using the 'Tumor ref.' reference (Smart-seq2-based) and conversely, the Melanoma dataset (Smart-seq2-based) was classified using the Ovarian dataset as a reference. The results, presented in **Fig. 2D**-F and **Fig. 3A** respectively, show a performance similar to that obtained within one platform. Taking the first of the two cross-platform classifications as an example (**Fig. 2F**) it is clear that the majority of cells (79.9%) that are not Tumor or Unknown retain the published labels. Of all the cells getting a different cell type

label most become Unassigned or Intermediate (87.3%) and this is especially true for the Unknown (80.0%) and Tumor cells (99.3%), in line with expectation. The robustness of the cross-platform classification is probably due to the use of rank-based similarities, implying that other combinations of scRNA-seq technologies will likely yield similar good results. This is further exemplified by accurate classification of a Drop-seq dataset[41] using a Chromium 10x Genomics dataset[28] (**Fig. 3C**).

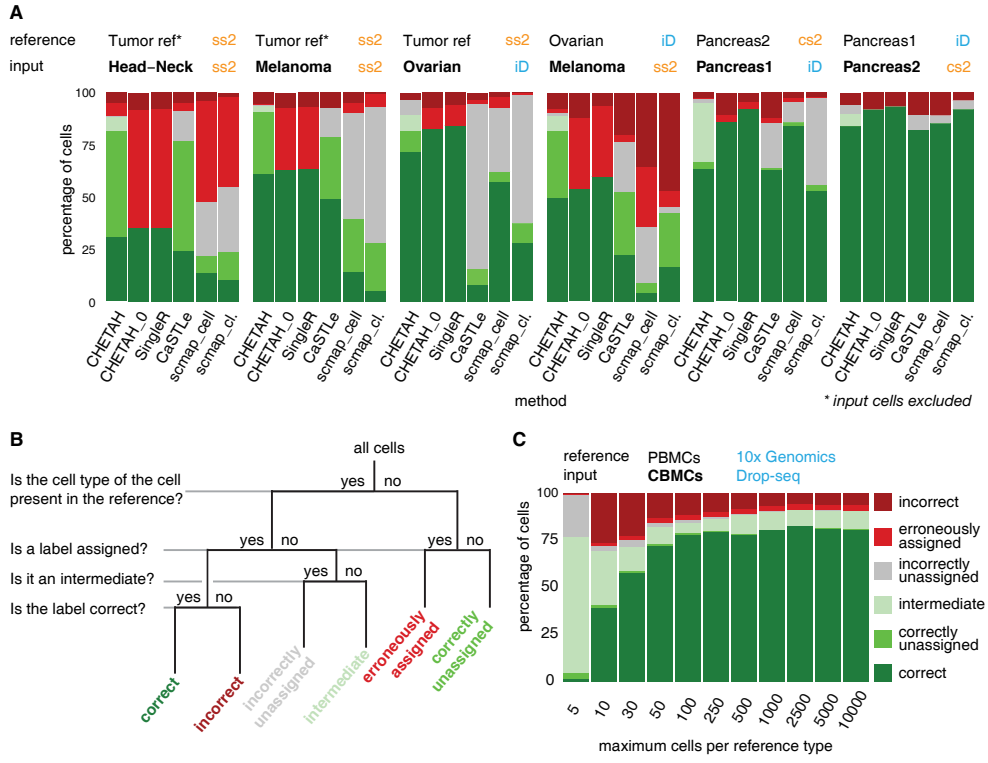


Figure 3. CHETAH compared with other methods (bottom labels), across six combinations of input and reference datasets (top labels, including the corresponding scRNA-seq platform: ss2: Smart-seq2; iD: inDrops; cs2: CEL-seq2. Microfluidics methods in blue, well-plate methods in orange). For scmap, both the 'cell' mode (scmap_cell) and 'cluster' mode (scmap_cl.) where evaluated. CHETAH was run with default settings, but also with a zero confidence score threshold (CHETAH_0), thus forcing it to classify all cells to a final type. **A)** Percentages of cells per classification result category as shown in B. **B)** Classification result categories used in A. **C)** The influence of the number of cells per reference cell type on CHETAH's classification performance was investigated as follows. The 7,830 cells of the (Drop-seq protocol) CITE-seq study[41], were classified with reference cells from the PBMC dataset[28], generated with the 10x Genomics platform. This dataset contained a total of 68,579 cells. The numbers on the y-axis are the number of (randomly sampled) cells per reference cell type taken to classify the input dataset. Classification results were divided into the six categories depicted in B. Besides investigating the influence of the number of cells in a reference type, this analysis also serves as an example of cross-platform performance, as well as an example using datasets with large numbers of cells. More details of the datasets used can be found in Table 1. Note that in the other analyses reported throughout, no limitation is placed on the number of cells per reference type.

Comparison with existing methods

The important challenge of cell type identification has recently also started to be addressed through the development of other automated approaches. CHETAH was therefore next compared to the state-of-the-art methods CaSTLe[42], scmap[31] (both versions, i.e. scmap_cell and scmap_cluster) and SingleR[32] by running these programs with standard settings on the Ovarian, Melanoma and Head-Neck datasets (**Fig. 3A**). To evaluate the performance also on non-tumor tissues, two pancreas datasets, Pancreas1[17] and Pancreas2[43] were included and mutually classified using the other as the reference. The ground truth for the classifications are the cell type labels from the original publications, but without the malignant cells from the tumor datasets. They are not part of the reference data and should therefore be considered an unknown cell type and should remain unassigned or intermediate.

To compare methods, two classes of input cells can be distinguished, namely [1] the cells that are of a type that is present in the reference and [2] cells for which no reference is available (**Fig. 3B**). For the first class it is meaningful to assess the correctness of the classification, because an optimal method should correctly identify all such cells. Those cell type inferences can therefore be correct or incorrect, corresponding to the true and false positives respectively (**Fig. 3B**). In addition, the categories intermediate or unassigned are allowed, to accommodate methods such as scmap and CHETAH that produce intermediate and unassigned calls. The second class of input cells, those of a type absent from the reference, should not be classified by an optimal cell type identification. These are therefore divided into correctly unassigned cells which can be considered true negatives, and their false positive counterparts, here called erroneously assigned, i.e. cells that were, but should not have been, classified.

In the cancer datasets, CHETAH generally outperforms other methods (**Fig. 3A**) in terms of combined true positives (correct assignments) and true negatives (cells correctly left unclassified). This is particularly important for studies of cancer since malignant cells are typically very patient specific and would almost always be misclassified by greedy methods. SingleR, having no classification cut-off, always classifies all cells to a final type, leading to a large number of erroneously assigned cells in cancer samples with many malignant cells (**Fig. 3**). For example, both the cancer-associated fibroblasts (CAFs) and malignant cells are all classified as CAFs by SingleR. In datasets containing many unknown cells such as the malignant cells in the cancer samples, such approaches would therefore require very careful post hoc inspection of the classification on a per cell or per cluster basis, an approach that automated methods are meant to obviate. The selective nature of CHETAH makes the analysis much more efficient. As anticipated, forcing CHETAH to become greedy and classify all cells by applying a confidence score threshold of 0, yields a performance almost identical to SingleR's (**Fig. 3**).

In contrast to the cancer datasets, the pancreas data are less complex, containing cell types with strong differential gene expression and few unknown cells. Note that a perfect classifier should leave none of the cells in the Pancreas2 dataset unidentified,

because all its cell types are represented in the Pancreas1 reference. The converse is not true because for some of the cell types no distinction is made in Pancreas1. This is one reason that all the methods perform better on Pancreas2 (**Fig. 3A**). An additional reason is the low expression of standard Pancreas markers in one of the donor samples included in the Pancreas1 dataset (**Fig. S8**). In the comparison on non-cancer datasets, CHETAH's forte of rarely classifying cells without sufficient resemblance to the reference cell types is diminished. This results in a performance similar to that of the other methods (**Fig. 3**). However, as is exemplified below, the inclusion of an intermediate assignment can have benefits for such datasets too.

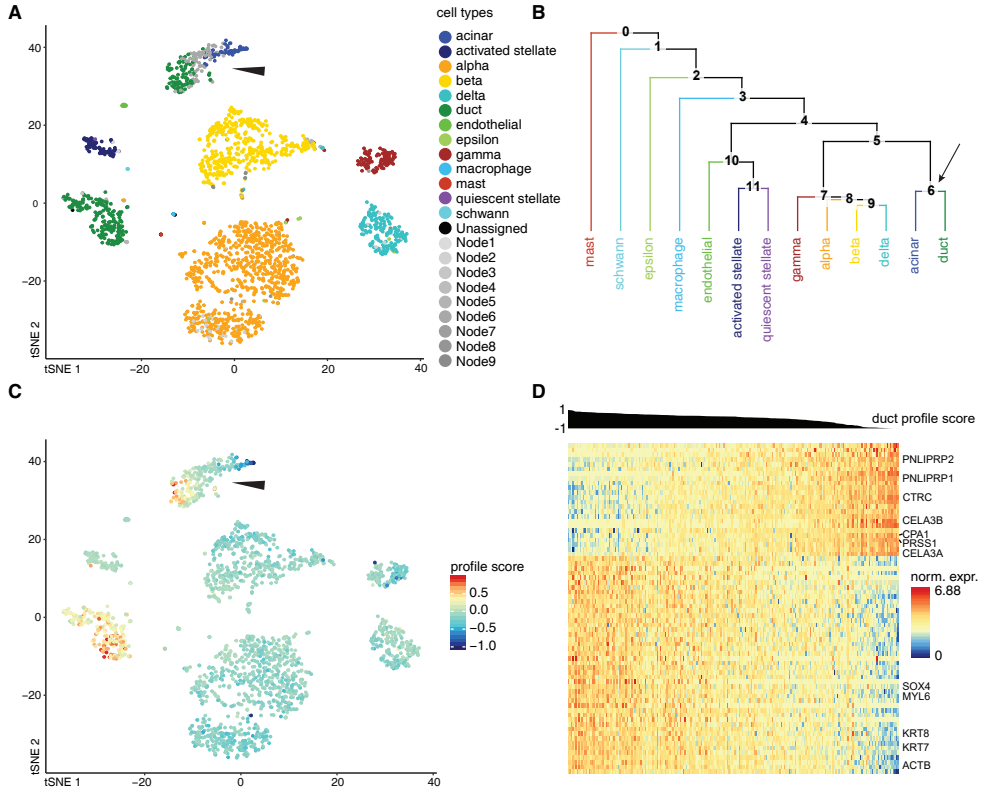


Figure 4. CHETAH identifies opposing gradients of duct and acinar cell marker genes in the Pancreas2 dataset[43]. **A)** t-SNE plot of the Pancreas2 dataset as classified by CHETAH, with colours representing the inferred cell types. The arrowhead indicates a population that was labelled as acinar cell in the publication, but is classified to a mixture of duct cell (blue), acinar cell (green) and intermediate Node 6 (grey) by CHETAH. **B)** The classification tree used for A, based on the Pancreas1 dataset. The arrow indicates the acinar/ductal intermediate node (Node 6) for which the profile score of duct cells is shown in C. **C)** As B, but with all cells coloured by the profile score for ductal cell in Node 6. The cells in the cluster of interest show a gradient of the profile score. **D)** Heatmap showing the normalised expression of the genes (rows) used by CHETAH to calculate the profile score plotted in C, for the cells (columns) in the cluster indicated by the arrowhead in panel A. Only genes (rows) having an absolute correlation > 0.5 with the profile score are shown. The cells are sorted by the duct cell profile score in Node 6 which is shown above the heatmap. Well-known acinar (top) and ductal marker genes (bottom) are labelled (see main text). For the heatmap with all genes annotated see Figure S7.

Intermediate types

In data from tumor samples the classification to an intermediate type suggests, by exclusion, that a cell is aberrant and therefore potentially malignant. The position in the classification tree, of the node of an intermediate type may shed further light on the biology of these cells. For example, in the Melanoma and Head-Neck datasets, 54% and 74% respectively of the malignant cells, classify to the node directly above endothelial. This suggests that the expression pattern of these cells shares characteristics with endothelial and fibroblast types (see **Fig. S3A,B** for the classification trees). Conversely, these cells display no affinity with the hematopoietic lineage, which is consistent with these tumors not being of hematopoietic origin. Classification to an intermediate type in combination with the position in the classification tree is therefore useful for analysis of cancer datasets.

Assignment to an intermediate cell type can also be useful in non-cancer datasets. This is demonstrated by two examples. In the Pancreas1 dataset, two kinds of stellate cells were originally identified, both of which are of mesenchymal origin[44]. PDGFRA and RGS5 were applied as marker genes for activated and quiescent stellate cells respectively. Pancreas2 only contains the more general label mesenchymal, and the corresponding cells only exhibit expression of PDGFRA but not RGS5 (**Fig. S6**), implying that these reference cells more closely resemble activated rather than quiescent stellate cells. When CHETAH classifies the Pancreas1 dataset using the more limited Pancreas2 reference data, it correctly identifies the Pancreas1 activated stellate cells as mesenchymal while leaving the quiescent stellate cells unassigned, or assigning them to the node directly above the endothelial and mesenchymal types (**Fig. S8B**), correctly determining that these cells are of a mesenchymal type not represented in the reference.

Acinar - duct cell gradient in pancreas data

Another useful consequence of allowing an intermediate type is exemplified in Fig. 4. Some cells in a cluster identified as acinar in the Pancreas2 publication are labelled ductal by CHETAH (**Fig. 4A**), while conversely the cluster called ductal in the Pancreas1 study is partly classified as acinar (**Fig. S8A**). The presence of these mixed acinar-ductal groups in both datasets suggests a shared underlying phenomenon. Acinar and ductal cells arise from the same progenitors and are closely related[45]. They are separated by only one node in CHETAH's classification tree (**Fig. 4B** and S8B), which is the intermediate type to which CHETAH assigns the remaining cells of these clusters. When visualising the profile score for duct cell in this intermediate node (arrows in **Fig. 4B** and **S8B**), a smooth gradient is clear in both clusters (**Fig. 4C** and **S8C**).

A heatmap of the expression of the genes most strongly (anti)correlating with this profile score shows the well-known cell type markers for these cell types (**Fig. 4D** and **S8D**). These cell type-specific markers again exhibit a gradient of decreasing ductal and increasing acinar expression. Among the negatively correlating genes are acinar markers like CPA1, PRSS1 and CTRC[46, 47] and among the positively correlating genes

are pancreas duct cell markers like KRT7 and KRT19[48]. A similar gradient in the expressions of genes having unusually large loadings in the first principal component of their ductal cell population has been reported previously[17]. This is a different manifestation of the fact that, for these cells, there is no dichotomy between acinar and ductal. Instead, the type of these cells is best described as lying on a continuum between acinar and ductal. The intermediate type assignment and profile score provide a direct and intuitive visualisation highlighting such cases and the utility of the approach taken by CHETAH.

Discussion

Classification of cell types in scRNA-seq data is an essential step that was by necessity initially performed manually[27,25,26,19]. Owing to the subjective and time-consuming nature of manual approaches, automated approaches have recently been developed[31, 32,42]. CHETAH has several features which work in its favour. Importantly, it compares input cell data with real, rather than imputed reference cell profiles. Moreover, besides using correlations, the classification decision is also based on a confidence score determined by the degree to which an input cell matches the expression variance embodied by the cumulative distribution function of the correlations to the reference cells. This facilitates the highly selective nature of CHETAH, underlying the ability to classify cells as specifically as the input and reference data allows, but without greedy over-classification, as controlled through the confidence score threshold. One consequence is the assignment to intermediate or unassigned cell types for input cells not present in the reference data. The assignment to an intermediate or unassigned type is essential to prevent overclassification and acts as an automated flag to more closely inspect such cells. The importance of this is evident both from the tumor datasets for which the method was initially devised, but also for non-cancer datasets as is also exemplified. In the tumor datasets analysed here practically all malignant cells were classified to intermediate types. Although genetic lesions such as copy number variations can be used to identify malignant cells[25], this does represent an additional step. Moreover, such aberrations are not necessarily present (as in many pediatric tumors[49, 50)] and/or may not be readily detectable. Automated highlighting of malignant tumor cells by CHETAH through classification as an intermediate or unassigned cell type is a significant improvement compared to blind misclassification. CHETAH's confidence threshold can be adjusted to the needs of the dataset at hand, making it a flexible tool for research. The method is made available as the R[33] package CHETAH in Bioconductor[34] release 3.9, useful for application in conjunction with tools such as SCENIC[9], Scater[10], Census[11], Monocle[12], Seurat[13], MERLoT[14] and CellBIC[15]. The CHETAH package additionally includes a Shiny[35] app for intuitive visualisation of the type labels, profile and confidence scores in a t-SNE[51] plot, as well as the inferred classification tree and expression heatmaps of discriminatory genes.

In the pancreas datasets, CHETAH uncovers a group of cells exhibiting a gradient of profiles between acinar and ductal, previously suggested to be centroacinar cells[17]. An alternative explanation is that these are acinar cells undergoing acinar-to-ductal

transdifferentiation or metaplasia (ADM)[47]. This is commonly seen in acinar cells that, like those in both pancreas studies discussed here, are cultured for several days[52] or subjected to stresses or injury[53]. Subtle phenomena such as the acinar-ductal gradient are easily overlooked by greedy methods and especially by (manual) methods that assign the same cell type to all cells of one apparent cluster. The accuracy of CHETAH is dependent on the availability of well-annotated reference datasets. It is firmly established that hierarchical trees derived from clustering gene expression data reflect many aspects of the underlying biology[54]. Here such trees are applied as a guide for classification only, without surmising accurate cellular taxonomies. Detailed hierarchical trees that reflect all aspects of cell types and cell states will obviously perform better for classification. Although the concordance between cell type identification based on cell surface markers and gene expression appears to be good[41], it is important to point out that gene expression is only one way of characterizing cells. The definition of cell types and the difference with cellular state are receiving renewed interest and scrutiny with the advent of quantitative single-cell techniques such as scRNA-seq (see e.g. ref 55). For the method presented here, the definition of cell type is pragmatic and can best be described as any group of cells annotated within a reference set as belonging together, and having sufficiently similar gene expression amongst themselves and sufficiently different gene expression with other types defined in the reference, so as to allow identification with high confidence. Classification of cells from diverse tissues, diseases and states will become easier with the increasing availability of well-annotated scRNA-seq datasets. Efforts like the Human Cell Atlas (HCA)[24] are aimed at generating scRNA-seq datasets for almost each (healthy) tissue and cell type. CHETAH's accurate handling of unknown cell types should prove useful in discovering novel cell types in such data. Conversely, the annotated HCA data would be very suitable as a reference for CHETAH.

Approaches for analysis of scRNA-seq data are being developed at a rapid pace. A recent addition is SuperCT[56] which incorporates supervised classification into a framework for cell-type classification. Although complementary in application scope (the reference dataset is fixed), we nevertheless compared accuracies, with CHETAH performing at a similar level of 92% concordance as analysed by the cross-validation method in the SuperCT study, albeit by necessity as tested on different datasets. CHETAH is not limited to the use of scRNA-seq and can likely be used with other quantitative single cell data such as those obtained using DNA accessibility[57, 58], chromatin state[59], methylome[60], epitope[41] or RNA velocity[61] sequencing methods, provided sufficiently rich reference data is available. Although the full range of single cell genome-wide approaches can be expected to increase further in the near future, the need for methods such as CHETAH that improve the ease and precision of the analysis of the resulting data is evident.

Software Availability

CHETAH is available at github.com/jdekanter/CHETAH and through Bioconductor[34]. All scripts that are needed to perform the analyses mentioned in this paper

and to create the t-SNE plots using Seurat[13] are deposited at github.com/jdekanter/CHETAH_paper_figures.

Acknowledgements

We thank members of the Holstege, Kemmeren and Molenaar groups at the Máxima Center for insightful discussions and comments.

Funding

The work presented in this paper is financially supported by the European Research Council (ERC) advanced grant 671174 and KIKA.

References

1. Islam,S., Kjällquist,U., Moliner,A., Zajac,P., Fan,J.-B., Lönnerberg,P. and Linnarsson,S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res., 21, 1160–1167.
2. Saadatpour,A., Lai,S., Guo,G. and Yuan,G.-C. (2015) Single-cell analysis in cancer genomics. Trends Genet. TIG, 31, 576–586.
3. Grün,D. and van Oudenaarden,A. (2015) Design and Analysis of Single-Cell Sequencing Experiments. Cell, 163, 799–810.
4. Lambrechts,D., Wauters,E., Boeckx,B., Aibar,S., Nittner,D., Burton,O., Bassez,A., Decaluwé,H., Pircher,A., Eynde,K.V. den, et al. (2018) Phenotype molding of stromal cells in the lung tumor microenvironment. Nat. Med., 24, 1277–1289.
5. Levitin,H.M., Yuan,J. and Sims,P.A. (2018) Single-Cell Transcriptomic Analysis of Tumor Heterogeneity. Trends Cancer, 4, 264–268.
6. Svensson,V., Vento-Tormo,R. and Teichmann,S.A. (2018) Exponential scaling of single-cell RNA-seq in the past decade. Nat. Protoc., 13, 599–604.
7. Moignard,V., Woodhouse,S., Haghverdi,L., Lilly,A.J., Tanaka,Y., Wilkinson,A.C., Buettner,F., Macaulay,I.C., Jawaid,W., Diamanti,E., et al. (2015) Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat. Biotechnol., 33, 269–276.
8. Lun,A.T.L., McCarthy,D.J. and Marioni,J.C. (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research, 5, 2122.
9. Aibar,S., González-Blas,C.B., Moerman,T., Huynh-Thu,V.A., Imrichova,H., Hulselmans,G., Rambow,F., Marine,J.-C., Geurts,P., Aerts,J., et al. (2017) SCENIC: single-cell regulatory network inference and clustering. Nat. Methods, 14, 1083–1086.
10. McCarthy,D.J., Campbell,K.R., Lun,A.T.L. and Wills,Q.F. (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinforma. Oxf. Engl., 33, 1179–1186.
11. Qiu,X., Hill,A., Packer,J., Lin,D., Ma,Y.-A. and Trapnell,C. (2017) Single-cell mRNA quantification and differential analysis with Census. Nat. Methods, 14, 309–315.
12. Qiu,X., Mao,Q., Tang,Y., Wang,L., Chawla,R., Pliner,H.A. and Trapnell,C. (2017) Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods, 14, 979–982.
13. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol., 36, 411–420.
14. Parra,R.G., Papadopoulos,N., Ahumada-Arranz,L., Kholtei,J.E., Mottelson,N., Horokhovskyi,Y., Treutlein,B. and Soeding,J. (2018) Reconstructing complex lineage trees from scRNA-seq data using MERLoT. bioRxiv, 10.1101/261768.
15. Kim,J., Stanescu,D.E. and Won,K.J. CellBIC: bimodality-based top-down clustering of single-cell RNA sequencing data reveals hierarchical structure of the cell type. Nucleic Acids Res., 10.1093/nar/gky698.
16. Stegle,O., Teichmann,S.A. and Marioni,J.C. (2015) Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet., 16, 133–145.
17. Baron,M., Veres,A., Wolock,S.L., Faust,A.L., Gaujoux,R., Vetere,A., Ryu,J.H., Wagner,B.K., Shen-Orr,S.S., Klein,A.M., et al. (2016) A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst., 3, 346-360.e4.
18. Zheng,C., Zheng,L., Yoo,J.-K., Guo,H., Zhang,Y., Guo,X., Kang,B., Hu,R., Huang,J.Y., Zhang,Q., et al. (2017)

Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell, 169, 1342-1356. e16.
19. Schelker,M., Feau,S., Du,J., Ranu,N., Klipp,E., MacBeath,G., Schoeberl,B. and Raue,A. (2017) Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat. Commun., 8, 2032.
20. Lane,K., Van Valen,D., DeFelice,M.M., Macklin,D.N., Kudo,T., Jaimovich,A., Carr,A., Meyer,T., Pe'er,D., Boutet,S.C., et al. (2017) Measuring Signaling and RNA-Seq in the Same Cell Links Gene Expression to Dynamic Patterns of NF-κB Activation. Cell Syst., 4, 458-469.e5.
21. Vladoiu,M.C., El-Hamamy,I., Donovan,L.K., Farooq,H., Holgado,B.L., Ramaswamy,V., Mack,S.C., Lee,J.J., Kumar,S., Przelicki,D., et al. (2018) Childhood cerebellar tumors mirror conserved fetal transcriptional programs. bioRxiv, 10.1101/350280.
22. Grün,D., Lyubimova,A., Kester,L., Wiebrands,K., Basak,O., Sasaki,N., Clevers,H. and van Oudenaarden,A. (2015) Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature, 525, 251–255.
23. Yang,J., Tanaka,Y., Seay,M., Li,Z., Jin,J., Garmire,L.X., Zhu,X., Taylor,A., Li,W., Euskirchen,G., et al. (2017) Single cell transcriptomics reveals unanticipated features of early hematopoietic precursors. Nucleic Acids Res., 45, 1281–1296.
24. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M., et al. (2017) The Human Cell Atlas. eLife, 6, e27041.
25. Tirosh,I., Izar,B., Prakadan,S.M., Wadsworth,M.H., Treacy,D., Trombetta,J.J., Rotem,A., Rodman,C., Lian,C., Murphy,G., et al. (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science, 352, 189–196.
26. Puram,S.V., Tirosh,I., Parikh,A.S., Patel,A.P., Yizhak,K., Gillespie,S., Rodman,C., Luo,C.L., Mroz,E.A., Emerick,K.S., et al. (2017) Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell, 171, 1611-1624.e24.
27. Patel,A.P., Tirosh,I., Trombetta,J.J., Shalek,A.K., Gillespie,S.M., Wakimoto,H., Cahill,D.P., Nahed,B.V., Curry,W.T., Martuza,R.L., et al. (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science, 344, 1396–1401.
28. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J., et al. (2017) Massively parallel digital transcriptional profiling of single cells. Nat. Commun., 8, ncomms14049.
29. Lyons,Y.A., Wu,S.Y., Overwijk,W.W., Baggerly,K.A. and Sood,A.K. (2017) Immune cell profiling in cancer: molecular approaches to cell-specific identification. Npj Precis. Oncol., 1, 26.
30. Han,X., Wang,R., Zhou,Y., Fei,L., Sun,H., Lai,S., Saadatpour,A., Zhou,Z., Chen,H., Ye,F., et al. (2018) Mapping the Mouse Cell Atlas by Microwell-Seq. Cell, 172, 1091-1107.e17.
31. Kiselev,V.Y., Yiu,A. and Hemberg,M. (2018) scmap: projection of single-cell RNA-seq data across data sets. Nat. Methods, 15, 359–362.
32. Aran,D., Looney,A.P., Liu,L., Wu,E., Fong,V., Hsu,A., Chak,S., Naikawadi,R.P., Wolters,P.J., Abate,A.R., et al. (2019) Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol., 10.1038/s41590-018-0276-y.
33. R Core Team (2018) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
34. Huber, W., Carey, J.V., Gentleman, R., Anders, S., Carlson, M., et al. (2015) Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods, 12, 115–121.
35. Chang,W., Cheng,J., Allaire,J.J., Xie,Y. and McPherson,J. (2018) shiny: Web Application Framework for R.
36. Yuan,H., Yan,M., Zhang,G., Liu,W., Deng,C., Liao,G., Xu,L., Luo,T., Yan,H., Long,Z., et al. (2019) CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res., 47, D900–D908.
37. Li,H., Courtois,E.T., Sengupta,D., Tan,Y., Chen,K.H., Goh,J.J.L., Kong,S.L., Chua,C., Hon,L.K., Tan,W.S., et al. (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet., 49, 708–718.
38. Chung,W., Eum,H.H., Lee,H.-O., Lee,K.-M., Lee,H.-B., Kim,K.-T., Ryu,H.S., Kim,S., Lee,J.E., Park,Y.H., et al. (2017) Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat. Commun., 8, 15081.
39. Picelli,S., Faridani,O.R., Björklund,Å.K., Winberg,G., Sagasser,S. and Sandberg,R. (2014) Full-length RNA-seq from single cells using Smart-seqNat. Protoc., 9, 171–181.
40. Zilionis,R., Nainys,J., Veres,A., Savova,V., Zemmour,D., Klein,A.M. and Mazutis,L. (2017) Single-cell barcoding and sequencing using droplet microfluidics. Nat. Protoc., 12, 44–73.
41. Stoeckius,M., Hafemeister,C., Stephenson,W., Houck-Loomis,B., Chattopadhyay,P.K., Swerdlow,H., Satija,R. and Smibert,P. (2017) Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods, 14, 865–868.
42. Lieberman,Y., Rokach,L. and Shay,T. (2018) CaSTLe – Classification of single cells by transfer learning:

Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. PLOS ONE, 13, e0205499.
43. Muraro,M.J., Dharmadhikari,G., Grün,D., Groen,N., Dielen,T., Jansen,E., van Gurp,L., Engelse,M.A., Carlotti,F., de Koning,E.J.P., et al. (2016) A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst., 3, 385-394.e3.
44. Erkan,M., Adler,G., Apte,M.V., Bachem,M.G., Buchholz,M., Detlefsen,S., Esposito,I., Friess,H., Gress,T.M., Habisch,H.-J., et al. (2012) StellaTUM: current consensus and discussion on pancreatic stellate cell research. Gut, 61, 172–178.
45. Reichert,M. and Rustgi,A.K. (2011) Pancreatic ductal cells in development, regeneration, and neoplasia. J. Clin. Invest., 121, 4572–4578.
46. Athwal,T., Huang,W., Mukherjee,R., Latawiec,D., Chvanov,M., Clarke,R., Smith,K., Campbell,F., Merriman,C., Criddle,D., et al. (2014) Expression of human cationic trypsinogen (PRSS1) in murine acinar cells promotes pancreatitis and apoptotic cell death. Cell Death Dis., 5, e1165.
47. Askan,G., Deshpande,V., Klimstra,D.S., Adsay,V., Sigel,C., Shia,J. and Basturk,O. (2016) Expression of Markers of Hepatocellular Differentiation in Pancreatic Acinar Cell Neoplasms A Potential Diagnostic Pitfall. Am. J. Clin. Pathol., 146, 163–169.
48. Bouwens,L. (1998) Cytokeratins and cell differentiation in the pancreas. J. Pathol., 184, 234–239.
49. Gröbner,S.N., Worst,B.C., Weischenfeldt,J., Buchhalter,I., Kleinheinz,K., Rudneva,V.A., Johann,P.D., Balasubramanian,G.P., Segura-Wang,M., Brabetz,S., et al. (2018) The landscape of genomic alterations across childhood cancers. Nature, 555, 321–327.
50. Ma,X., Liu,Y., Liu,Y., Alexandrov,L.B., Edmonson,M.N., Gawad,C., Zhou,X., Li,Y., Rusch,M.C., Easton,J., et al. (2018) Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature, 555, 371–376.
51. Maaten,L. van der and Hinton,G. (2008) Visualizing Data using t-SNE. J. Mach. Learn. Res., 9, 2579–2605.
52. Houbracken,I., de Waele,E., Lardon,J., Ling,Z., Heimberg,H., Rooman,I. and Bouwens,L. (2011) Lineage Tracing Evidence for Transdifferentiation of Acinar to Duct Cells and Plasticity of Human Pancreas. Gastroenterology, 141, 731-741.e4.
53. Strobel,O., Dor,Y., Alsina,J., Stirman,A., Lauwers,G., Trainor,A., Castillo,C.F., Warshaw,A.L. and Thayer,S.P. (2007) In Vivo Lineage Tracing Defines the Role of Acinar-to-Ductal Transdifferentiation in Inflammatory Ductal Metaplasia. Gastroenterology, 133, 1999–2009.
54. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci., 95, 14863–14868.
55. What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? (2017) Cell Syst., 4, 255–259.
56. Xie,P., Gao,M., Wang,C., Zhang,J., Noel,P., Yang,C., Von Hoff,D., Han,H., Zhang,M.Q. and Lin,W. SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. Nucleic Acids Res., 10.1093/nar/gkz116.
57. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. Nature, 523, 486–490.
58. Lake,B.B., Chen,S., Sos,B.C., Fan,J., Kaeser,G.E., Yung,Y.C., Duong,T.E., Gao,D., Chun,J., Kharchenko,P.V., et al. (2018) Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol., 36, 70–80.
59. Rotem,A., Ram,O., Shoresh,N., Sperling,R.A., Goren,A., Weitz,D.A. and Bernstein,B.E. (2015) Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat. Biotechnol., 33, 1165–1172.
60. Hou,Y., Guo,H., Cao,C., Li,X., Hu,B., Zhu,P., Wu,X., Wen,L., Tang,F., Huang,Y., et al. (2016) Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell Res., 26, 304–319.
61. Manno,G.L., Soldatov,R., Hochgerner,H., Zeisel,A., Petukhov,V., Kastriti,M., Lonnerberg,P., Furlan,A., Fan,J., Liu,Z., et al. (2017) RNA velocity in single cells. bioRxiv, 10.1101/206052.

Supplemenatry Material

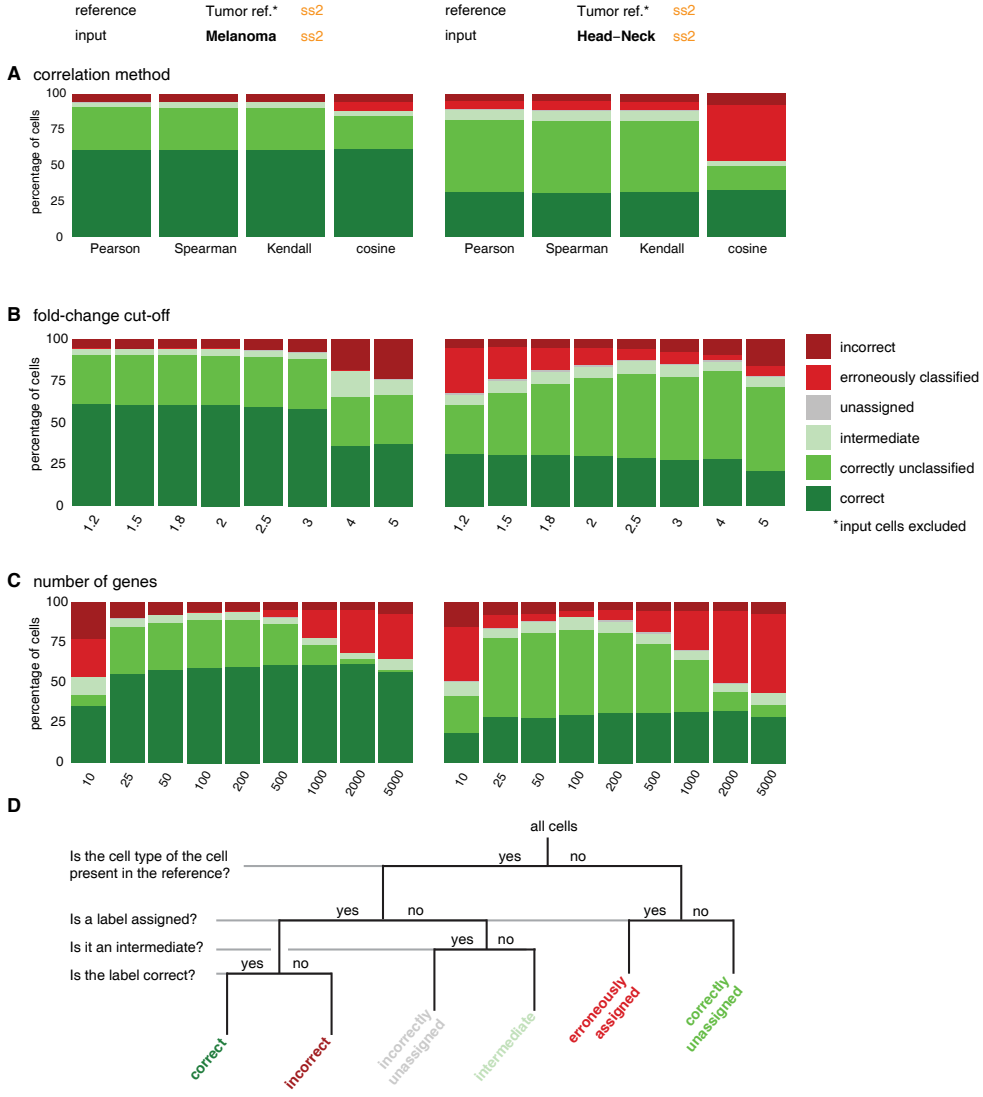


Figure S1. Influence of three of CHETAH's selectable parameters on its classification performance, for two different data sets and references (for a description of datasets see Table 1; classification types as in panel D. **A)** Dependence on correlation method for four correlation methods. **B)** Dependence on feature selection method where the genes changing more than the indicated fold-change are selected as most discriminatory gene set. **C)** Dependence on feature selection method where the indicated number of with the largest absolute difference are selected. **D)** Classification result categories used in **A-C**.

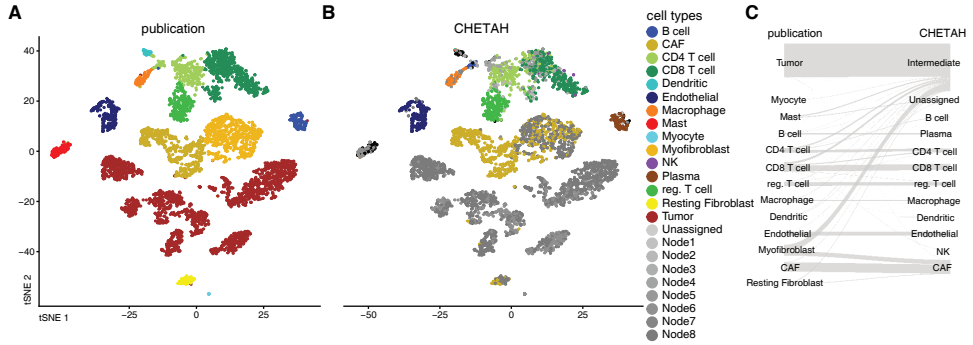


Figure S2. Classification of the Head-Neck tumor data (Head-Neck dataset: Puram et al., 2017, reference 26 in main text). Details as in main Figure 2.

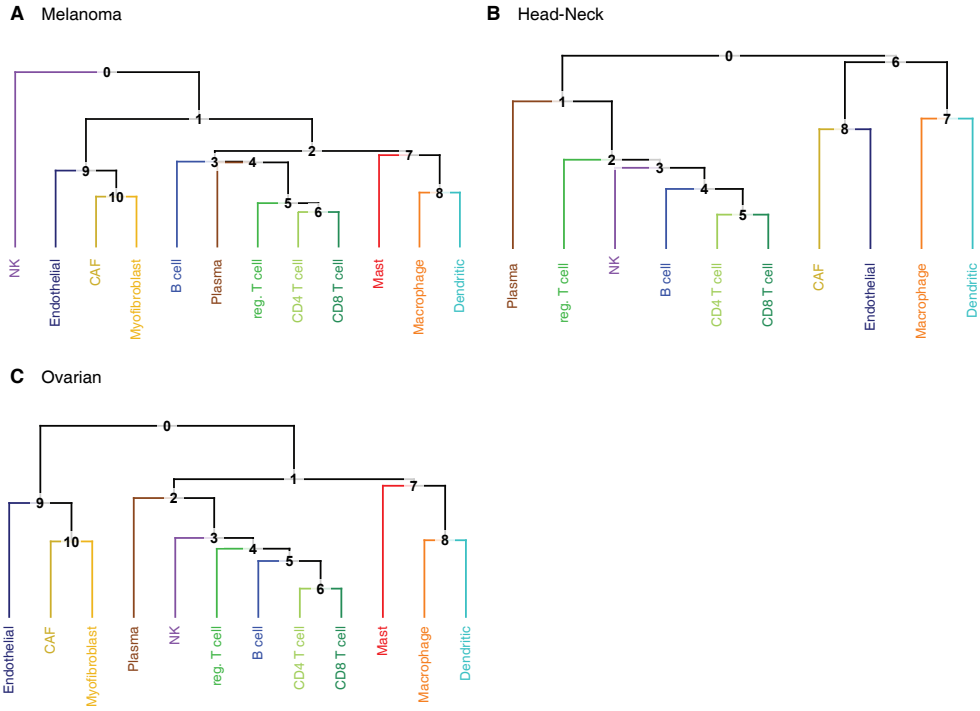


Figure S3. The classification trees used for the tumor datasets. Note that the trees are similar and reflect known biology. Small differences can arise due to a limited number of reference cells available (e.g. just 14 NK reference cells in the Melanoma classification tree). In the Melanoma and Head-Neck datasets, respectively 54% and 74% of the cells labeled as malignant cells classify to the node above *cancer associated fibroblasts* and *endothelial*.

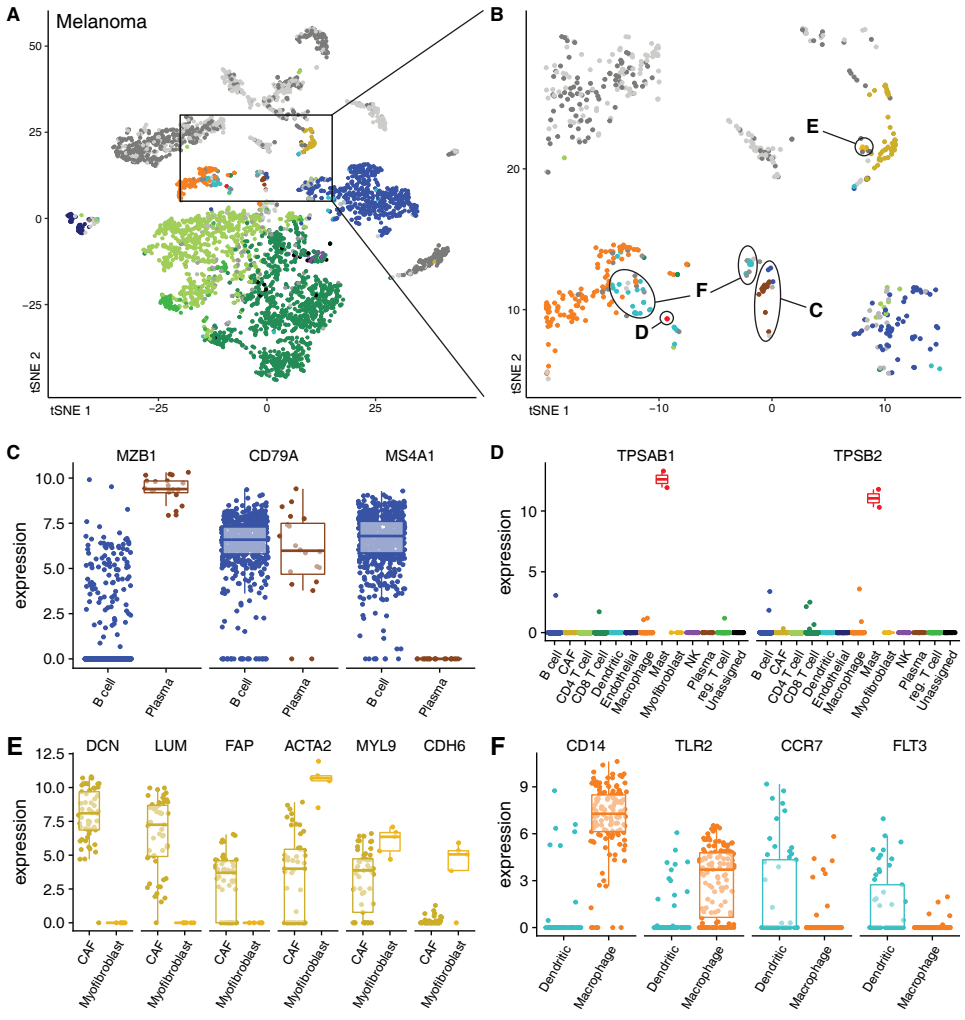


Figure S4. When CHETAH classification differs from the one assigned in the Melanoma study (Figure 2A), marker gene expression suggests CHETAH's classification is more plausible. t-SNE plots of **(A)** the Melanoma dataset and **(B)** a zoom-in of this dataset. Details are as in main Figure 2. The arrows point to the discrepantly classified populations with letters C-F indicating the panel in which the marker gene expression of this population is plotted. **(C-F)**: For each cluster shown in panel B: boxplots of marker gene expression (number of tran–scripts) of the marker genes (shown on top) for cells of the conflicting cell types (CHETAH classification shown at the bottom). Colours are the same as in panels A,B. **C)** Most of the cells previously classified as B cells in the Melanoma dataset are probably naive B cells, but some (encircled C in panel B) are more likely plasma cells. Most of these cells express B cell marker CD79A. Most plasma cells express activation marker MZB1 but lack expression of CD20/MS4A1 which goes down upon activation. The naive B cells express these two genes in opposite manner. **D)** Two previously unclassified cells are probably mast cells. These are the only cells that highly express mast markers TPSAB1 and TPSB2. **E)** Previously unclassified cells in the cancer-associated fibroblasts (CAFs) cluster are probably myofibroblasts. These cells express none of the CAF markers (DCN, LUM, FAP), while they express higher levels of actin and myosin genes compared to the CAFs. **F)** Previously unclassified cells are probably dendritic cells. Few of these cells express macrophage markers (CD14, TLR2), while they are enriched for dendritic markers (CCR7, FLT3).

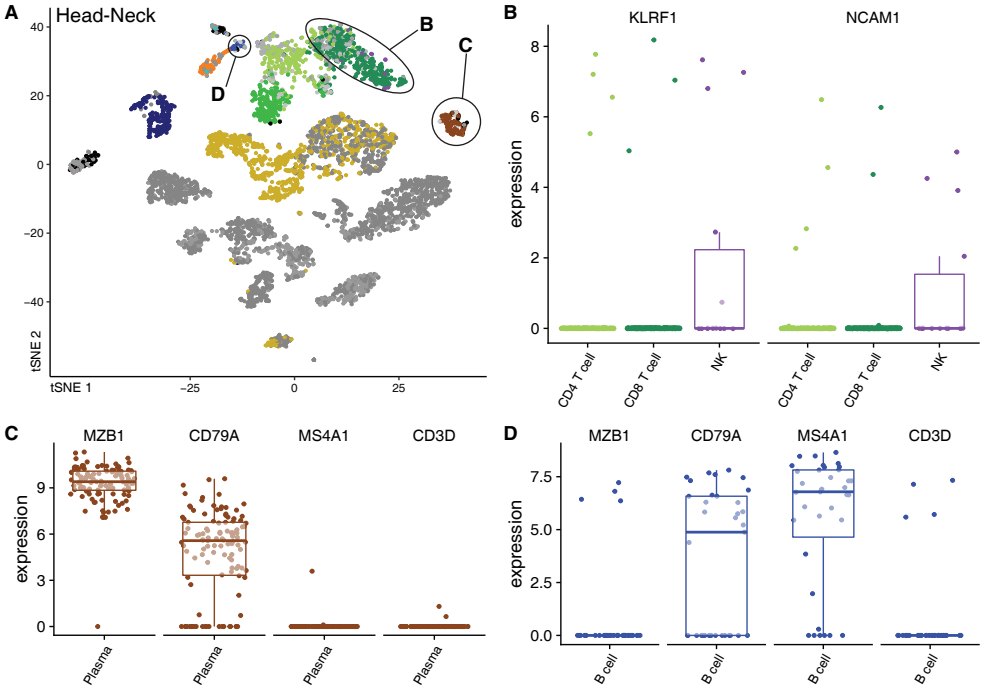


Figure S5. When CHETAH classification differs from the one assigned in the Head-Neck study (Figure 1B, S1), marker gene expression suggests CHETAH's classification is more plausible. Details as in Figure S3. **A)** t-SNE plots of the Head-Neck dataset coloured by CHETAH classification. **B)** Cells previously classified as CD8 T-cells (they cluster together strongly) are probably NK cells. These cells are enriched for NK marker genes KLRF1 and NCAM1. **C)** Cells previously identified as B-cells are probably active plasma cells. Most cells express B-cell marker CD79A as well as activation marker MZB1 while lacking expression of CD20/MS4A1. **D)** Cells previously classified as T-cells are probably naive B cells. Few of these cells express T-cell marker CD3D, while most express B cell marker CD79A. These cells show no expression of plasma cell marker MZB1, but do express CD20

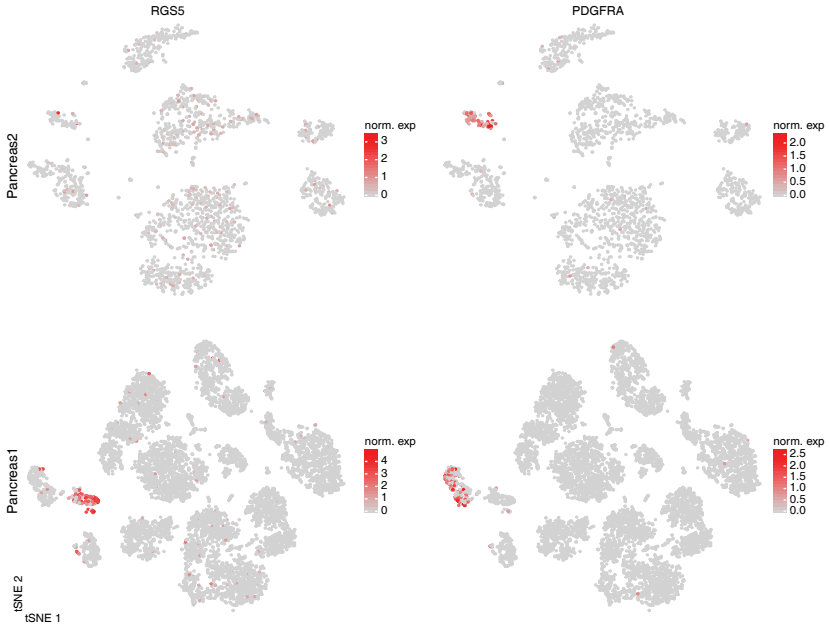


Figure S6. t-SNE plots coloured by the normalized expression of the gene indicated. Rows of the panels: the data classified, as indicated; columns: gene whose expression is shown. In the Pancreas1 publication the RGS5 gene was used as a marker for quiescent stellate cells whereas PDGFRA was used as a marker for activated stellate cells. The Pancreas2 dataset only has expression of PDGFRA, not of RGS5.

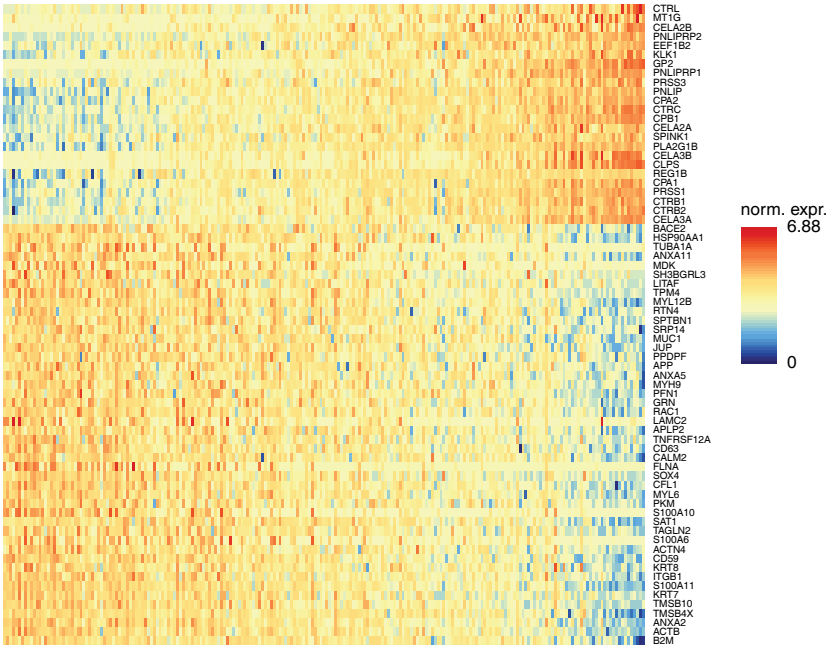


Figure S7. The heatmap shown in main Figure 4D, but with all genes labeled.

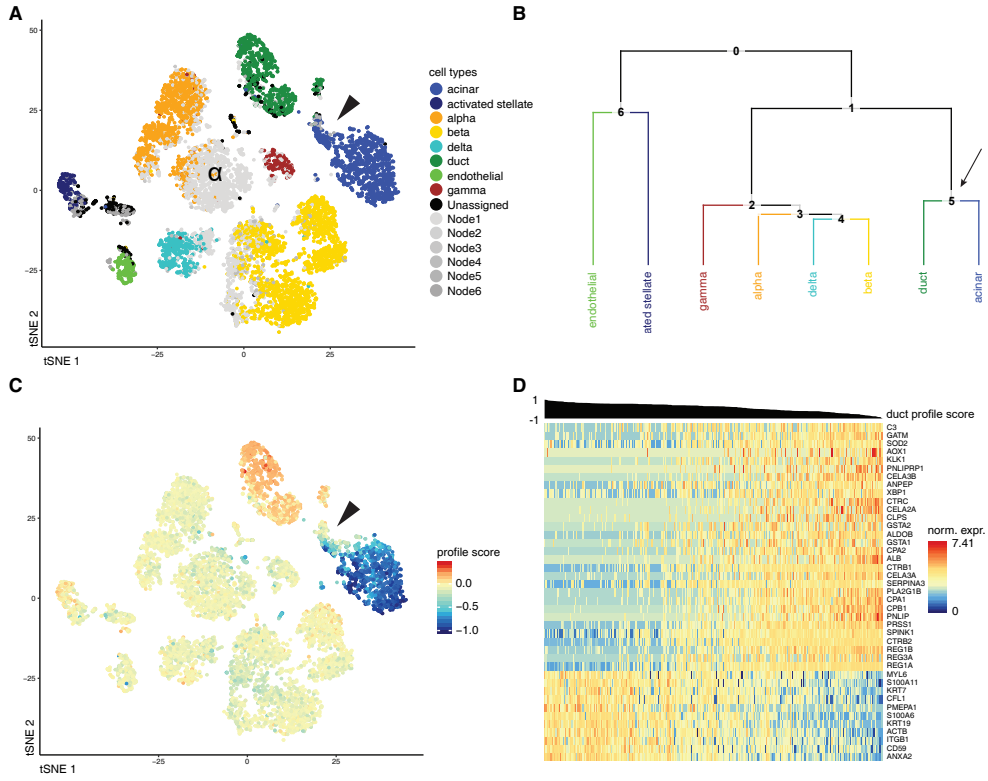


Figure S8. CHETAH identifies opposing gradients of duct and acinar cell marker genes in the Pancreas1 dataset (Baron et al., 2016; reference 10 in main text). **A)** t-SNE plot of the Pancreas1 dataset as classified by CHETAH, with colours representing the inferred cell types. The arrowhead indicates a population that was labeled as duct cells in the publication, but is classified to a mixture of acinar cells and intermediate type 6 by CHETAH. **B)** The classification tree used for A, based on the Pancreas2 dataset. The arrow indicates the acinar/ductal intermediate node (Node 5) for which the profile score of duct cells is shown in C. **C)** As B, but with cells coloured by the profile score for ductal cell in Node 5. The cells in the cluster of interest show a gradient of the profile score. **D)** Heatmap showing the normalized expression of genes used by CHETAH to calculate this profile score. Only the genes that correlate, or anti-correlate more than 0.5 with this profile score in these populations are shown. Rows are genes, columns are cells. The columns are sorted by the duct cell profile score which is shown above the heatmap. For the heatmap with all genes annotated see Figure S7. Note: both CHETAH and scmap have difficulty classifying the alpha cells of one of the donors (as identified in the original publication, here indicated by 'α' in panel A as they have an unusually low expression of the relevant markers.


7

Single-cell RNA sequencing reveals heterogeneous T cell inhibition in pediatric Hodgkin Lymphoma

Jurrian K. de Kanter[1,2,*], Alexander Steemers[1,2,*], Ravian L. van Ineveld[1,2], Niels Groenen[1,2], Marijn A. Scheijde-Vermeulen[1], Liset Westera[1], Auke Beishuizen[1], Frank C.P. Holstege[1], Anne C. Rios[1,2], Arianne M. Brandsma[1], Thanasis Margaritis[1], Ruben van Boxtel[1,2,*], Friederike Meyer-Wentrup[1,*]

[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 to this work

Manuscript under review

Abstract

Pediatric classic Hodgkin lymphoma (cHL) patients have a high survival rate but suffer from severe long-term side effects induced by chemo- and radiotherapy. cHL tumors are characterized by the low fraction (0.1-10%) of malignant Hodgkin Reed-Sternberg (HRS) cells in the tumor. The HRS cells depend on the surrounding immune cells for survival and growth. This dependence is leveraged by current treatments that target the PD-1/PD-L1 axis in cHL tumors. The development of more targeted therapies that are specific for the tumor and are therefore less toxic for healthy tissue compared to conventional chemotherapy could improve the quality of life of pediatric cHL survivors. Here, we applied single-cell RNA-sequencing (scRNA-seq) on isolated HRS cells and the immune cells from the same cHL tumors. This allowed us to identify genes of cell surface proteins that are consistently overexpressed in HRS cells and can potentially be used as targets for antibody-drug conjugates or CAR T cells. Finally, we identify potential interactions by which HRS cells inhibit T cells, among which the Galectin-1/CD69 and HLA-DRA/LAG3 interactions. However, high levels of inter-patient heterogeneity of the interaction strength were observed. RNAscope was used to validate the enrichment of CD69 and LAG3 expression on T cells near HRS cells but indicated large variability of the interaction strength with the corresponding ligands between patients and between tumor tissue regions. In conclusion, this study identifies new potential therapeutic targets for cHL and highlights the importance of studying heterogeneity when identifying therapy targets.

Introduction

Classical Hodgkin lymphoma (cHL) accounts for 10-15% of all lymphoma cases and represents the most commonly diagnosed lymphoma subtype in adolescents and young adults (AYAs)[1]. Combined-modality treatment regimens composed of multiagent chemotherapy and involved-site radiation therapy have greatly improved cHL survival, with current cure rates exceeding 90%[1]. However, 10–30% of adult cHL patients and 10% of pediatric cHL patients have refractory or recurrent disease[2–4]. Despite the administration of high-dose chemotherapy supported by autologous hematopoietic stem cell transplantation (ASCT), adult refractory/relapsed (R/R) patients have a poor prognosis[5]. For pediatric R/R patients, the prognosis is better[6], but long-term toxicity of the treatment is a significant problem. A recent study has shown that pediatric HL survivors had 100 excess deaths per 10,000 person-years 25 years post-diagnosis and nearly 400 excess deaths ≥40 years from diagnosis, making HL the cancer with the second-highest long-term excess mortality after medulloblastoma[7]. These excess deaths are primarily attributed to treatment-related secondary malignant neoplasms and symptomatic cardiac/pulmonary toxicities[7]. Hence, there is high demand for novel and innovative treatment approaches that target the tumor more specifically and have reduced side effects while preserving or improving clinical efficacy.

The cellular ecosystem of cHL is unique as it consists of rare malignant Hodgkin Reed–Sternberg (HRS) cells which typically represent 0.1–10% of all cells in the tumor tissue and are surrounded by a dense immune microenvironment consisting of mostly lymphocytes, myeloid cells, and fibroblasts[8]. The CD30-positive HRS cells are believed to be derived from pro-apoptotic germinal center B cells as they have rearranged (non-functional) immunoglobulin genes, gone through somatic hypermutation, and lost the expression of B cell lineage markers such as CD19, CD79a, and immunoglobulin gene transcripts[9–11]. The following lines of evidence suggest that the rich immune infiltrate in cHL creates an immunosuppressive tumor microenvironment (TME) that is essential for supporting HRS cell survival and growth. First, HRS cells are tightly adhered to surrounding T cells, a phenomenon termed rosetting, which likely is essential for HRS cell survival[12]. Additionally, HRS cells do not survive in immunodeficient mice nor grow as solitary cells *in vitro*, and establishing HRS-derived cell lines has been proven difficult[13,14]. Given that HRS cells depend on complex interactions with different immune cells, breaking or interfering with these interactions represents a promising treatment strategy.

Currently, two immune checkpoint inhibitors have been FDA-approved for R/R cHL patients, namely nivolumab and pembrolizumab, both of which target the PD-1/PD-L1 signaling axis[15,16]. Of note, pembrolizumab is the only immunotherapy currently approved for pediatric cHL patients. Although the use of these PD-1 inhibitors in combination with standard chemotherapy regimens has led to a significant improvement in clinical outcome, a large portion of patients still relapses, highlighting the dire need for the development of alternative therapies, for example, those that interfere with interactions between HRS and immune cells that are essential for HRS cell survival[17].

Single-cell RNA sequencing (scRNA-seq) provides an opportunity to describe the TME in detail through precise molecular profiling of individual cells while simultaneously predicting tumor-immune cell interactions[18]. Previous studies applying scRNA-seq to cHL samples have already yielded novel insights into the cHL TME. For example, Aoki et al. identified a $LAG3^+$ regulatory T cell-like subpopulation that contributes to the immunosuppressive phenotype of cHL[19]. The same group later characterized a unique $CD4^+PD\text{-}1^+CXCL13^+$ T follicular helper cell-like subset in lymphocyte-rich cHL that surrounds HRS cells. The presence of this T cell population is associated with poor clinical outcome[20]. Furthermore, the transcription factors TOX and TOX2 were identified as key regulators of exhaustion in previously reported rosette-forming $CD4^+CD26^-$ T-cell populations[21]. Finally, dendritic cells, monocytes, and macrophages were found to be enriched in the close vicinity of HRS cells, all expressing immunoregulatory checkpoints including PD-L1, TIM-3, and the tryptophan-catabolizing protein IDO[22]. These discoveries have significantly improved our understanding of the pathogenic mechanisms that are active in the cHL microenvironment. However, there are several important limitations in the

published scRNA-seq studies, including the lack of HRS cells detected in the samples, as well as the under-representation of pediatric patients included in the cohorts.

Here, we performed flow cytometry-based cell enrichment combined with plate-based scRNA-seq to specifically capture HRS cells and simultaneously the TME cells of pediatric cHL samples. We used these data to show that the presence and strength of HRS-immune cell interactions are highly variable between patients. The interaction between HRS cells with LAG3$^+$ CD8 T cells and CD69$^+$ T cells was predicted based on scRNA-seq data and validated by RNAscope. In addition, we identified HRS cell genes of surface proteins, other than CD30, that were expressed by the majority of HRS cells but few healthy tissues and thus pose potential new therapy targets.

Results

Cell sorting and data processing

Two pediatric cHL lymph nodes were dissociated into single cells, sorted into 384-well plates, and processed using the SORT-seq protocol[23]. Events with high levels of side scatter (SSC+) were sorted into two columns of each plate to enrich for HRS cells. HRS cells could only be identified in the scRNA-seq data of one of the two samples, which had a higher-than-average (~5%) HRS content in the tumor based on diagnostic CD30 immunohistochemistry staining. Therefore, the sorting strategy was adjusted to enrich for HRS cells using five antibodies based on previous literature[24,25] (SSC$^+$CD20$^-$CD95$^+$CD15$^+$CD30$^+$CD40$^+$, **Fig. S1**). In addition, SSC$^+$CD20$^-$ cells were sorted to capture any other potential subset of the HRS cells. Seven additional pediatric cHL lymph nodes were processed with this panel. For six out of these seven samples, the HRS cell gate captured 0.01-0.04% of all the live singlets in each sample, while for the last sample (PB16107), this was more than 10 times higher (0.5%). For the last patient, more cells that expressed all HRS markers could therefore be sorted. Finally, 3 non-malignant reactive lymph nodes (RLN) were included in this study as controls. The majority of cells in RLN were CD3$^+$ and therefore SSC$^+$ (non-lymphocytes) and CD20$^+$ (B cells) cells were enriched in these samples. In total, 9594 wells were sorted and sequenced. 5710 wells passed QC and were included in the final data set. See **Table S1** for the number and sorting strategy of events per patient. The nine cHL patients were between 9 and 17 years old at the time of diagnosis, tumors were of stage II or III and had a supraclavicular or cervical localization, and one patient (PB09287) had EBV-positive HRS cells (**Table 1**).

Plate based scRNA-seq captures HRS cells

Classification of the cell types was guided by automated cell type identifiers CHETAH[26] and SingleR[27] (**Fig. S2A,B**). B and T lymphocytes and myeloid cells were identified (**Fig. 1A,B**). Each of these cell types was processed and clustered separately, and subtypes were identified based on canonical subtype markers (**Fig. S2C-E**). Exhausted and cycling CD8 T cells were identified, as well as exhausted, cycling, and naive CD4 T cells and T follicular helper cells (TFH, **Fig. 1A**). Germinal center

B cells (GC-B), and naive and plasma B cells were detected, as were monocytes and macrophages, together with different subsets of dendritic cells (DC). Of note, some cell types, like cycling CD8 and CD4 T cells, were highly enriched in the cells sorted on specific flow cytometry characteristics and were only present in small numbers in the unbiased live cells.

patient	diagnosis	sub type	stage	EBV	location	age	sex	scRNA-seq	HRS gate (%)	RNAscope
PB24752	cHL	ns	NA	neg	NA	15	male	yes	-	no
PB19568	cHL	ns	II A	neg	cervical	16	female	yes	0.01	no
PB09287	cHL	mc	II A	pos	cervical	10	male	yes	0.01	no
PB31727	cHL	nos	III B	neg	cervical	12	male	yes	0.04	no
PB26217	cHL	ns	III A	neg	cervical	17	male	yes	-	yes
PB16107	cHL	ns	II AE	neg	supraclavicular	16	female	yes	0.48	no
PB11473	cHL	ns	II A	neg	supraclavicular	9	female	yes	0.01	yes
PB05088	cHL	ns	II BE	neg	cervical	14	male	yes	0.03	no
PB09908	cHL	ns	III AE	neg	cervical	15	male	yes	0.01	no
PB06422	cHL	nos	III B	NA	NA	15	male	no	-	yes
PB27302	cHL	ns	VI BE	NA	NA	16	male	no	-	yes
PB25394	RLN	-	-	-	armpit	8	male	yes	-	no
PB32331	RLN	-	-	-	cervical	13	male	Yes	-	no
PB32684	RLN	-	-	-	NA	15	male	yes	-	no

Table 1. Sample information and clinical information of the patients included in this study.
RLN = reactive lymph node, cHL = classic Hodgkin Lymphoma. ns = nodular sclerosis, mc = mixed cellularity, nos = not otherwise specified.

In addition, a few clusters were detected in the scRNA-seq data that separated from the other cell types. These clusters only contained data from cHL samples (n = 8), and the cells of most individual patients clustered separately. As opposed to the immune cells in the data set, 75% of these cells were classified by CHETAH as an intermediate cell type, indicating they were not of any cell type in the CHETAH cancer TME reference dataset (**Fig. S2A**). These cells followed expression patterns found on HRS cells by immunohistochemistry (**Fig. S3A,B**). Indeed, 93% of these were sorted by the HRS cell gate, indicating that the HRS markers were expressed both on RNA and protein levels in these cells (**Fig. 1C**). Furthermore, when inferring chromosomal copy numbers using inferCNV[28], large chromosomal gains and losses were detected, which are characteristic of HRS cells (**Fig. 1D**). We observed recurrent amplifications

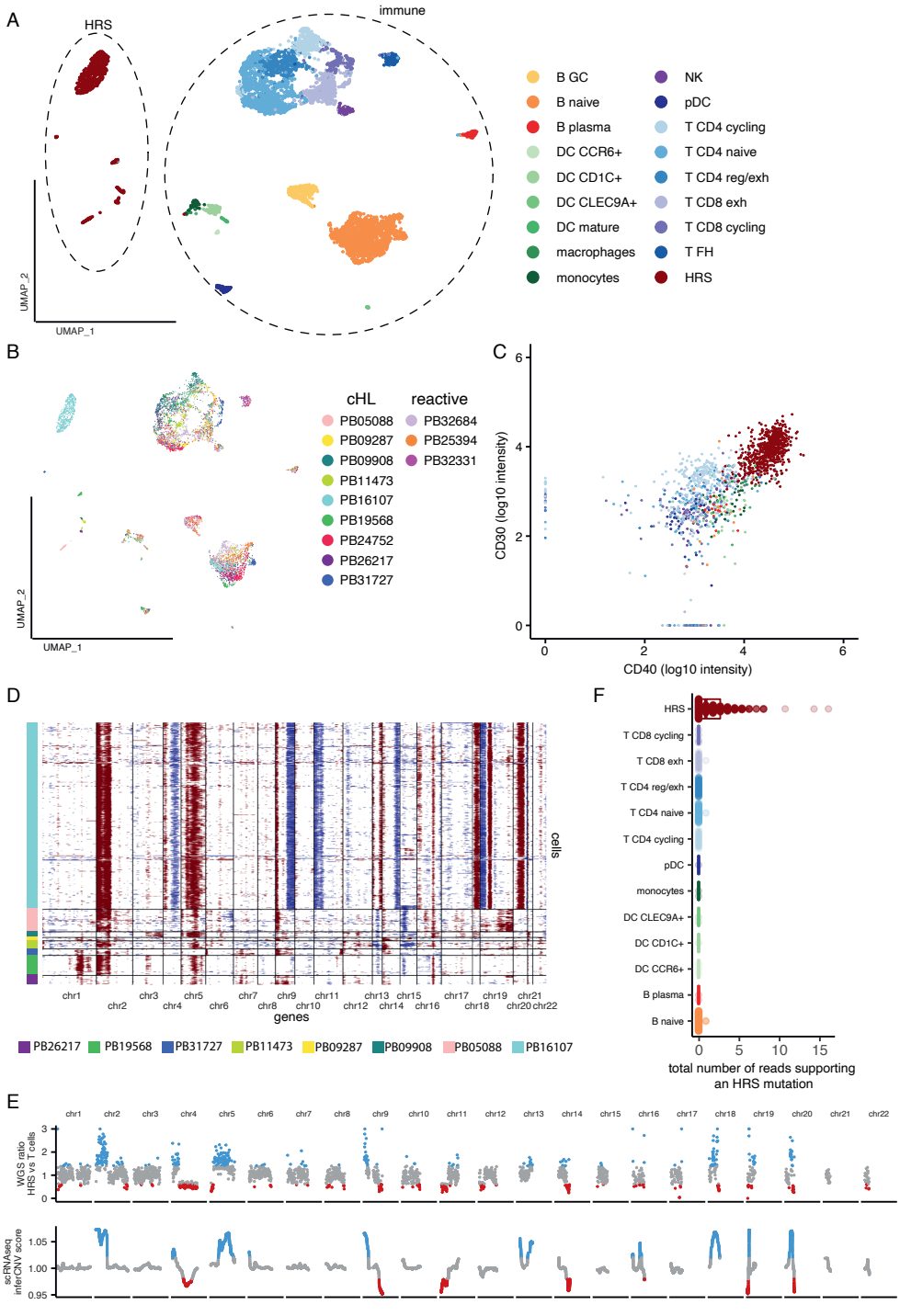


Figure 1. HRS cells were captured by plate-based scRNA-seq

Legend on the next page

A) UMAP plot of all cells from cHL and reactive lymph nodes (RLN) labeled by cell type. **B)** UMAP plot colored by patient. **C)** FACS intensities of the cells sorted with CD30 and CD40 antibodies labeled according to the scRNA-seq cell types. **D)** Copy number variation (CNV) plot of the HRS cells shown in (A). Each row is a cell, each column is a gene. **E)** Normalized copy number plots of patient PB16107 based on HRS cell WGS data (top) and HRS cell scRNA-seq (bottom). **F)** The number of reads in each cell of patient PB16107 that supported a mutation found in the WGS data of HRS cells from the same patient.

of chromosomes 2p, 5, and 9p (n = 4), and recurrent deletions of chromosomes 13 (n = 3), which is in line with previous studies[10,29]. To validate that these cells were indeed HRS cells, whole genome sequencing (WGS) was applied to DNA extracted from 3,500 sorted HRS cells of patient PB16107 using low-input whole genome amplification. The CNV pattern that was inferred from the scRNA-seq HRS cluster of this patient was highly similar to WGS-based CNVs (**Fig. 1E**). To further validate that these cells were HRS cells, 22 WGS-based single-base substitutions (SBS) were identified that had sufficient coverage in the scRNA-seq data. 99.5% of the 980 unique reads that supported HRS cell SBS were from cells in the HRS cluster (**Fig. 1F**). Together, these results validate the HRS cell identity.

As HRS cells have rosetting T cells that can remain attached throughout the cell sorting procedure, we investigated the expression of T cell genes in the HRS clusters. Although some events in the HRS clusters did express a few T cell markers, only CD4 was expressed to similar levels as CD4 T cells, indicating that if present, the T cells contributed relatively few transcripts compared to the HRS cells (**Fig. S3C,D**).

Defining a core set of HRS marker genes

As described above, genes that are expressed by most HRS cells in most patients could pose novel targets for cHL treatment. In addition, highly specific HRS markers can potentially simplify the identification of HRS cells, e.g., by decreasing the number of antibodies needed for flow cytometry purification. To overcome the potential biases of the single dataset from our center and to identify targets that are relevant for pediatric and adult cHL, we performed differential expression analysis (DEA) between HRS cells and healthy B cells in both the scRNA-seq and two microarray datasets of micro-dissected HRS cells[13,30]. 837 genes were consistently overexpressed in HRS cells in the 3 datasets and were together termed the "HRS-core" set (**Fig. 2A, B**). As expected, the HRS-core set was depleted for genes located on the recurrently deleted chromosome 13 (p=0.03) and seemed enriched for genes located on the recurrently amplified chromosome 2, albeit not significantly (p=0.08). We validated the HRS-core set using bulk RNA-seq of cHL lymph nodes and RLN obtained from the diagnostics department. The expression of most HRS-core genes correlated well with the expression of HRS marker *TNFRSF8* (CD30) in cHL bulk RNA-seq, underlining their specificity for HRS cells (**Fig. S4A, B**). Of the HRS-core set, 122 genes were also differentially expressed between cHL and RLN samples. Of these, 74 were also uniquely overexpressed in bulk cHL compared to other B cell lymphomas.

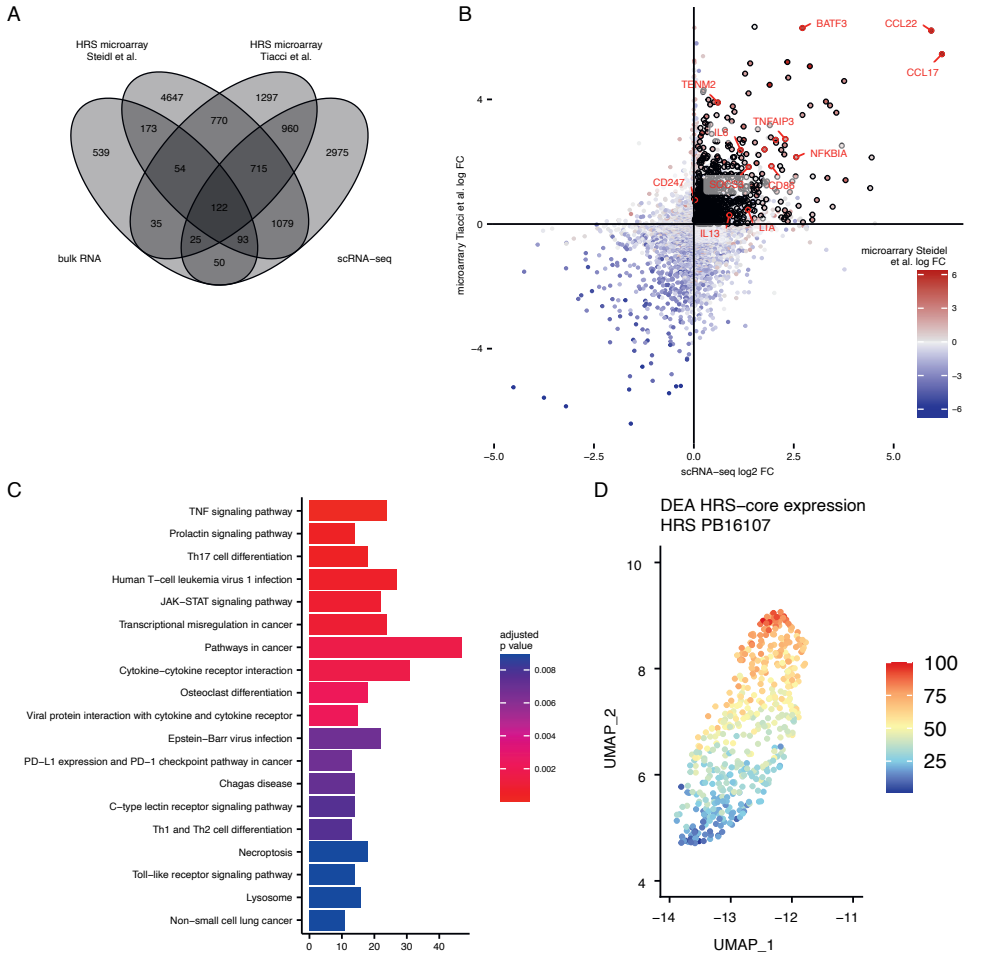


Figure 2. HRS core-genes identification

A) A Venn diagram of HRS markers as identified in four datasets, HRS cell microarray data from Steidl et al.30 and Tiacci et al.[13], the scRNA-seq data presented here, and bulk RNA-seq data of cHL and reactive lymph nodes. The genes that overlapped between the two microarray datasets and the scRNA-seq dataset were termed "HRS-core" genes. **B)** The differential expression of HRS markers in HRS cells compared to normal B cells in the scRNA-seq data compared to the Tiacci et al. microarray data. Each point is a gene. Points are colored according to fold change in expression in the Steidl et al. microarray data of HRS cells compared to healthy B cells. **C)** KEGG-pathway enrichment of the HRS-core genes. D) An aggregate score of the expression of differentially expressed HRS-core genes in HRS cells of patient PB16107.

The HRS-core genes were enriched for gene ontology (GO) terms involved in pathways that were previously reported to be active in HRS cells such as extrinsic apoptotic signaling, positive regulation of leukocyte cell-cell adhesion, positive regulation of T cell activation, and mononuclear cell migration (**Fig. S4C**). In addition, the HRS cells were enriched for KEGG pathways JAK-STAT (*JAK3, STAT1/5A/5B, SOCS1/2/3*), TNF (*TNF, TNFAIP3*), EBV infection (*CD44, E2F1*), and PD-L1 checkpoint (*CD274,*

BATF, BATF3), all of which have previously been identified to be active in HRS cells (**Fig. 3C**)[13,30,31]. We also identified genes encoding for well-described signaling molecules that play central roles in cHL pathology, such as *IL6, IL13, IL15, CCL17* (TARC), *CCL22* (MDC), and *LTA* (TNF-β)[32–35]. Corresponding receptor genes *IL13RA1* and *IL15RA* were also part of the HRS-core set, suggesting that IL-13 and IL-15 might be the interleukins that are most commonly involved in autocrine HRS cell signaling. These interleukins might thus play a central role in HRS cell survival and therefore pose potential targets for therapy. Genes encoding other well-described interleukins such as IL4, IL5, IL8, and IL10, and their receptor[31,36,37], were not consistently overexpressed in HRS cells in all three datasets. Furthermore, genes that are normally only expressed in other tissues were identified, for example, the *TENM2/3, ADCY1, BRINP2* (nervous system), and *DHRS2* (bladder). These genes are likely expressed due to the chromosomal rearrangements within the HRS cells[38].

To identify potentially therapeutically targetable HRS genes, genes were selected with a high predicted likelihood of being present on the plasma membrane, being expressed in only a few healthy tissues, and being present on the HRS cells of all (but one) cHL patients in the scRNA-seq. This resulted in 10 genes including the canonical HRS marker *TNFRSF8* (CD30) which is targeted by the clinically applied brentuximab vedotin (anti-CD30)[16,39,40]. In addition, interleukin gene *IL6* was identified, as well as *TNF*, testis-specific lipoprotein-receptor *LRP8*, lipoprotein *APOL1*, keratinocyte-specific *PERP*, and the well-described Epstein-Barr Virus Induced 3 (*EBI3*), which heterodimerizes with p28 to form IL27. IL6, TNF, and IL27 can be membrane-bound, but they would more likely be useful targets for unconjugated antibodies that block their binding to target cells. This approach is most likely to be successful if these genes are proven to be essential in the cHL TME, and risk severe side effects as such antibodies could potentially affect the entire immune system. Based on their function, most of the other genes encode for proteins that are unlikely to play key roles in HRS signalling pathways. However, they could potentially be used as therapeutic targets for antibody-drug conjugates or CAR T cells, or in flow cytometry to identify HRS cells.

HRS cell heterogeneity

Investigating the intra-patient heterogeneity of HRS cells was not possible for most samples, as the number of HRS cells was too low (9-53 cells). For PB16107, 497 HRS cells were captured, which were therefore processed separately. Interestingly, these HRS cells formed a continuum, with 35% of differentially expressed genes between the two ends of this continuum overlapping with the HRS-core set (*CLL17, CCL22, TNF, LTA, NFKBIA, IL6, IL13*). This was independent of total UMI counts or cell cycle phase (**Fig. 2D**). This finding suggests that the general HRS-core expression "program" can have varying levels of activity within the HRS cells of a single patient.

To further analyze diversity, we investigated the inferred CNV profiles of PB16107 and found only 1 minor subclone, which had a loss of chromosome 15 (**Fig. 1C**). No subcluster of cells could be found by PCA or t-SNE that had a lower expression of genes positioned on chromosome 15, suggesting that chromosomal instability did not drive the highest levels of gene expression heterogeneity in the HRS cells.

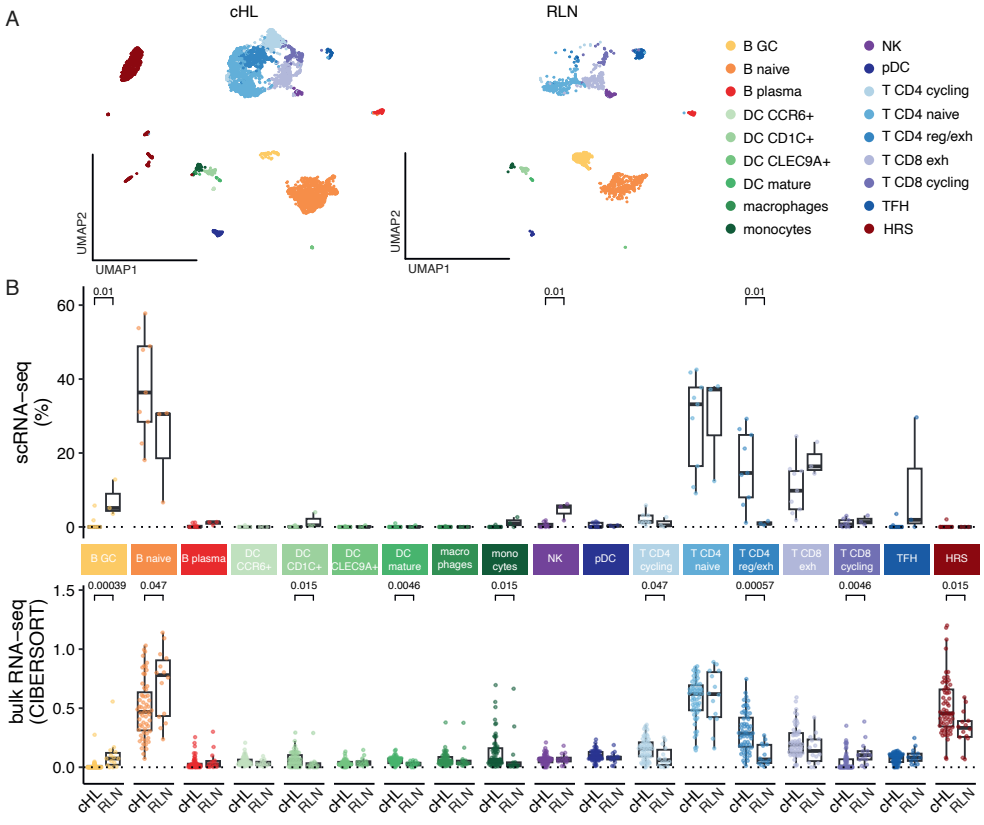


Figure 3. The immune cell composition of the cHL microenvironment differs from reactive lymph nodes

A) UMAP plots of cells of cHL lymph nodes and reactive lymph nodes (RLN) labeled by cell type. **B)** A quantification of the percentage of cell types shown in (A) per sample. Each dot represents a sample. Here and in all other figures, 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). P-values were calculated using the differential composition analysis of DCATS41 and fdr-corrected. **C)** The estimated frequency of cell types in bulk RNA-seq data of cHL lymph nodes and RLN as estimated by CIBERSORT. P-values were calculated by the Wilcoxon-test and fdr-corrected.

HRS cells inhibit T cells by a variety of interactions

To study cell-cell interactions in cHL, it is important to first identify which cell types are enriched and depleted in the cHL TME compared to normal lymph nodes. For this analysis, only unbiased live cells were used that were not enriched for any marker in flow cytometry. Exhausted CD4 T cells were overrepresented in cHL compared to

RLN (relative ratio 16.6, p=0.01, **Fig. 3A, B**). In contrast, GC-B cells, the B cell type that are abundant in normal germinal centers, was present at lower frequencies in cHL lymph nodes, as were NK cells (relative ratio, 0.11 and 0.12, p=0.01). Plasma cells and TFH were also depleted, although not significantly, due to their low numbers in both cHL and controls.

To validate these results in a larger number of samples, CIBERSORTx[42] was used for deconvolution of bulk RNA-seq of 59 cHL and 14 RLN obtained from the diagnostics department, using the scRNA-seq data as a reference (**Fig. 3C**). This analysis validated that the exhausted CD4 T cell was the most enriched cell type in cHL compared to reactive lymph nodes (ratio 3.0, p=0.0005), and that GC-B cells were most depleted in cHL samples (ratio 9.4, p=0.0003). In addition, cycling CD8 T cells were depleted (p=0.0046) and cycling CD4 T cells were enriched (p=0.047) in bulk RNA-seq cHL samples, although this difference was small in the scRNA-seq data (1% compared to 2% of all cells in both cases). In addition, a higher number of myeloid cells was found in bulk cHL compared to bulk RLN (p=0.015), while this was not the case in scRNA-seq. These cells likely have a lower survival during the freeze-thawing and single-cell sorting procedure and are therefore underrepresented. In addition, while in the scRNA-seq naive B cells were more common in cHL lymph nodes, in deconvolution, these cells were more abundant in RLN. These results validated that germinal center B cells are depleted from cHL tissue, and that exhausted CD4 T cells are the most enriched type. It seems therefore likely that HRS cells interact most with exhausted CD4 T cells.

Our data provides a unique opportunity to investigate *in vivo* interactions between HRS cells and the TME on a per-patient basis. Interaction scores were calculated by grouping cell types into broader categories and multiplying the expression of a ligand/receptor in a TME cell type of one patient with the expression of the corresponding receptor/ligand in the HRS cells of the same patient (**Fig. 4A**). Most interactions that we identified were only observed in one or a few patients (**Fig. S5**) and 49% of the interactions that were found to be active in cHL had a similar or stronger activity in RLN. In CD4 and CD8 T cells, B cells, and DCs an enrichment could be observed for interactions with HRS cells that were present in all eight investigated cHL lymph nodes. Of these, HRS had the highest number of consistent interactions with CD4 T cells.

The high number of exhausted CD4 T cells in cHL, and the high number of predicted interactions between HRS and CD4 T cells suggest an important role for this cell type in HRS survival. Indeed, some of the strongest interactions in cHL had a role in the inhibition/exhaustion of T and NK cells. Compared to RLN, CD4 T cells expressed more *CTLA4*. *LAG3* was overexpressed in the CD8 T cells in our data set

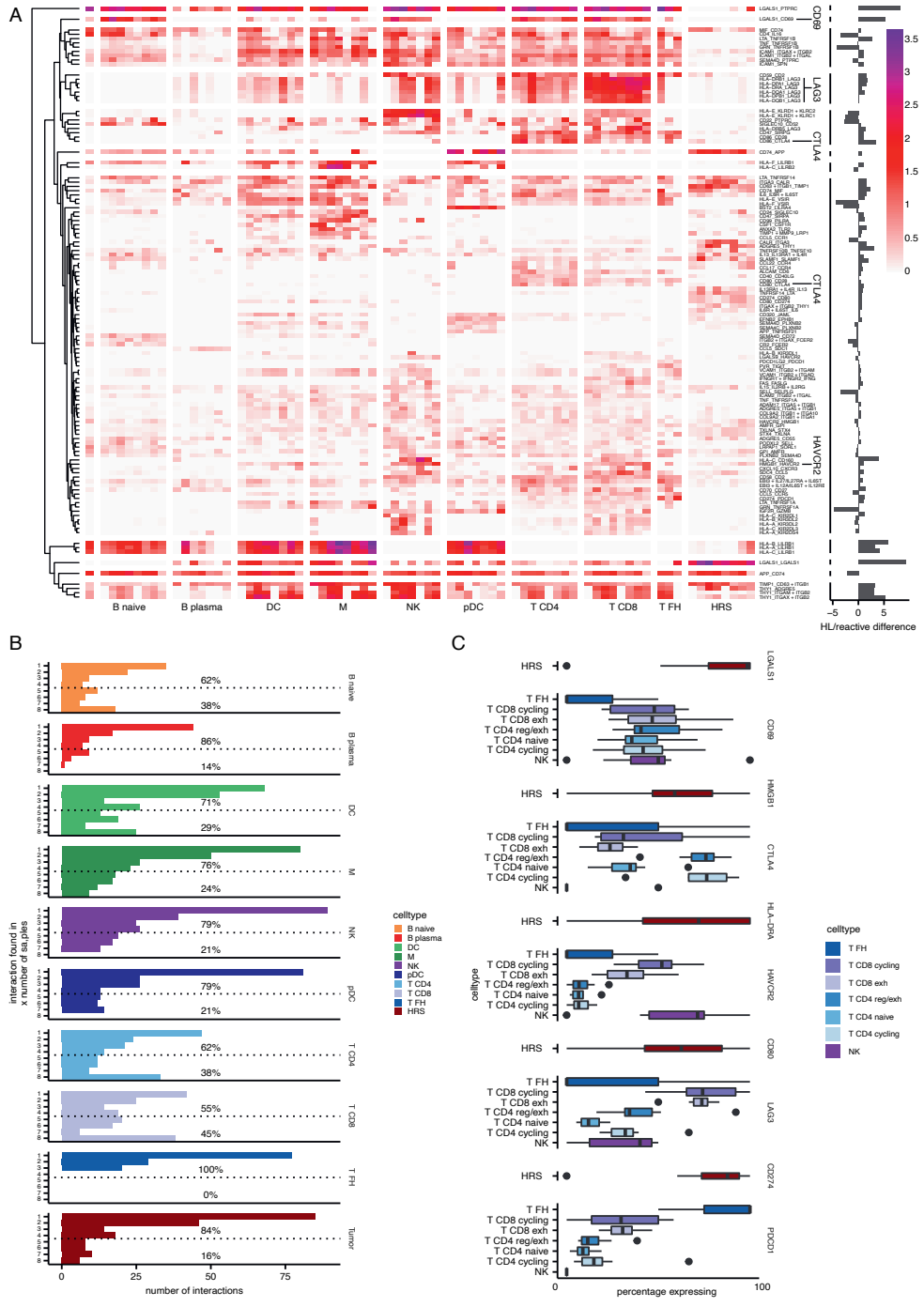


Figure 4. Potential interactions between HRS cells and TME cells were identified.
A) The strength of HRS cell interactions with other cells in the microenvironment. Each block is the interaction strength with a cell type in one patient. Each interaction is annotated with [HRS cell ligand]_[immune cell receptor]. *The legend continues on the next page.*

Only interactions are shown that are present between HRS cells and a single immune cell type in 3 or more samples. The difference in the maximal interaction strength is indicated on the right side of the plot. **B)** The number of samples in which an interaction was present between HRS cells and the indicated cell type. **C)** The percentage of cells expressing inhibitory receptors on T or *NK* cells per sample. For each receptor, the percentage of HRS cells expressing the corresponding ligand is depicted. Each dot is an outlier.

(**Fig. 4C**), while in adults, *LAG3* is mostly expressed in CD4 T cells[19]. In addition, NK cells expressed more *HAVCR2* (TIM3), and myeloid cells expressed the inhibitory immunoglobulin-like gene *LILRB1* (CD85J). Ligands for these receptors were expressed in HRS cells (*CD80, HLA-II, HMGB1, HLA-I,* **Fig. 4C**). *PD-1L* is highly expressed in HRS cells due to 9p24.1 alterations[43], which is confirmed in our data. However, the PD-1L/PD1 interaction was not present. The only subtype of T cells that expressed PD1 in our dataset was the TFH cell (**Fig. 4C**). TFH cells were depleted in cHL lymph nodes in our data, which might explain the previous observation that cHL tumors are depleted of PD1-expressing T cells[44]. Finally, Galectin-1 (*LGALS1*) is highly and specifically expressed in HRS cells of all samples. Galectin-1 can induce T cell exhaustion via the CD69 receptor, which is expressed on a subset of cHL T cells[45]. Targeting this interaction has been proposed as a general method to enhance T cell anti-tumor immunity in cancers[46] and might thus be of interest for developing strategies to treat cHL using targeted approaches. However, the interaction scores varied greatly between patients. For example, the highest interaction score for CTLA4-CD80 in CD4 T cells was 14 times higher than the lowest score. The ratio between the highest and lowest interaction score was 7 for CD69-LGALS1, and the ratio was one of the lowest for LAG3-HLA-DRA at 1.8, indicating this is one of the most common and consistent interactions in pediatric cHL and thus likely important for HRS survival.

Inter-patient spatial heterogeneity of commonly detected interactions

To validate the presence of the inhibitory interactions between HRS and T cells and to investigate the inter-patient heterogeneity of these interactions, RNAscope imaging was performed (**Fig. 5A, S6**). For 4 patients, the interaction between HLA-DRA$^+$ HRS cells and LAG3$^+$ CD8 T cells was studied (probes for *HLA-DRA, LAG3, CD8,* and *TNFRSF8,* refered to as *CD30*). In a separate panel the interaction between LGALS1$^+$ HRS cells with CD69$^+$ T cells was studied (probes for *LGALS1, CD3E, CD69, TNFRSF8/ CD30*). Two of the samples were selected from the patients of the scRNA-seq cohort, two were from cHL patients outside of our scRNA-seq cohort (**Table 1**). RNAscope was applied on entire slices of the cHL lymph nodes. For each patient, between 4 and 8 representative regions with varying expression of the different markers were selected for further inspection. Interestingly, the total *LAG3* expression per region correlated with *CD8A* expression, as did *CD69* expression with *CD3E* (R2>=0.7 in each patient, **Fig. 5A,B**). Second, *CD8A* was more often co-expressed with *LAG3* in the same cells compared to background (**Fig. 5C**). *CD3E* was co-expressed with *CD69*, but to a lower extent. In summary, the RNAscope data confirms the frequent

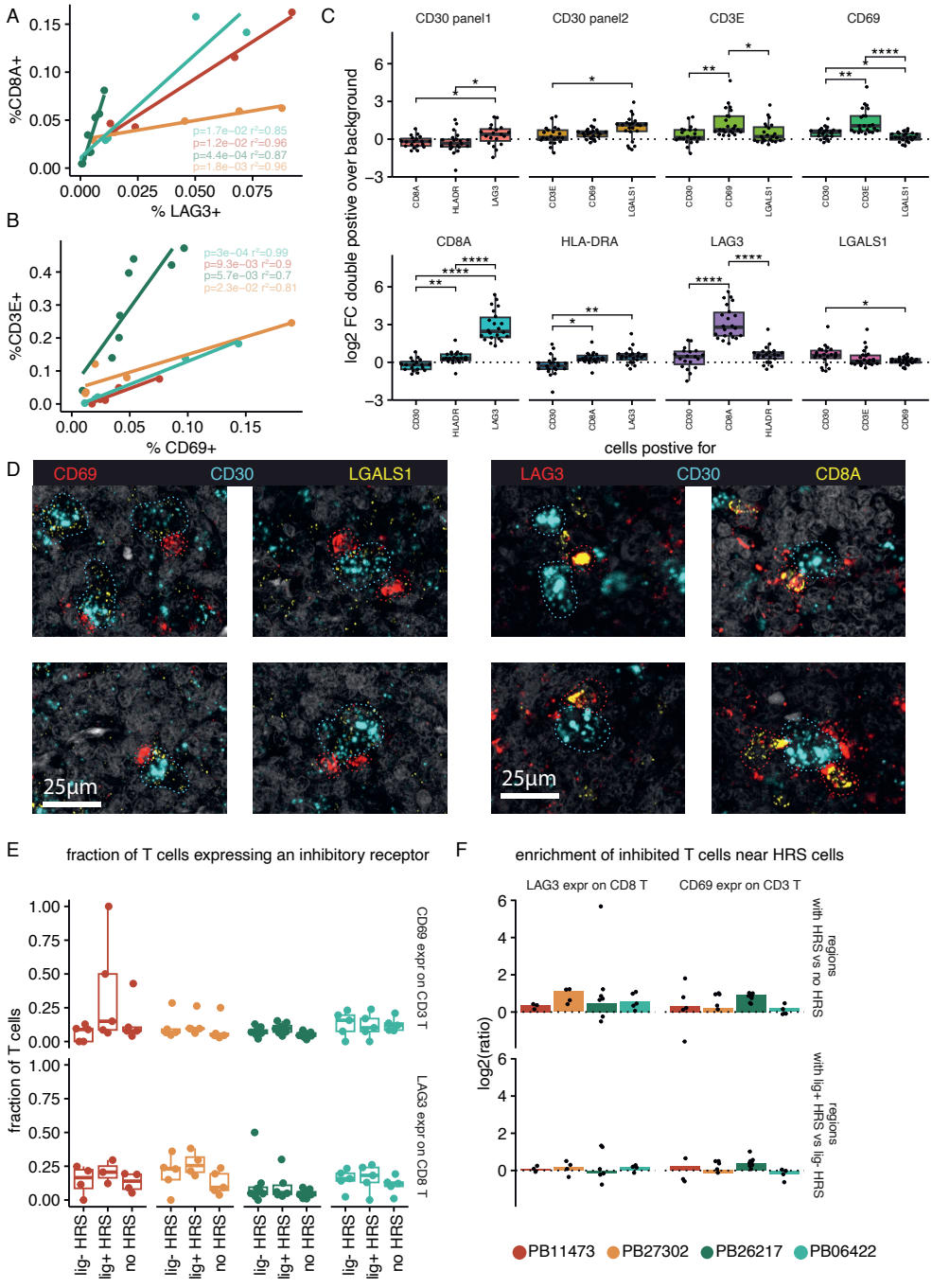


Figure 5. Spatial assessment of LAG3⁺CD8A⁺ and CD69⁺CD3E⁺ T cells with HRS cells
A) Examples of RNAscope images of HRS cells (CD30) with T cells in close proximity. On the left side images of RNAscope panel 1 of patient PB26217 are depicted. On the right, images of panel 2 of patient PB27302 are depicted. *The legend continues on the next page.*

B) RNAscope co-expression of *CD8A* and *LAG3* across regions of cHL lymph nodes. Regions were separated into 51px blocks. The fraction of blocks positive for each marker is depicted. Each dot is a region of one lymph node. **C)** The same as B, but for *CD69* and *CD3E* expression. **D)** Co-expression of genes in 51px blocks (7.2 μm). The fold change of block positive for a gene (on the x axis) that is also positive for a second gene (above each plot) compared to the background expression of that second gene in the other blocks. A value above 0 means co-expression is more often observed than expected by chance. Each dot is a single region in a sample. **E)** The fraction of T cells (as defined by *CD3E* or *CD8A* expression) that express the inhibitory receptor genes *LAG3* and *CD69*. T cells were separated into blocks surrounding HRS cells that did not express the inhibitory ligand ("lig- HRS"), blocks surrounding HRS cells that expressed the inhibitory ligand ("lig$^+$ HRS"), and regions not adjacent to any HRS cell ("no HRS"). Ligands are *LGALS1* for the *CD69* receptor and *HLA-DRA* for the *LAG3* receptor. **F)** Enrichments of the data shown in E. Top: T cells expressing the inhibitory receptor gene near HRS cells ("lig- HRS"/"lig$^+$ HRS" in E) compared to T cells not near HRS cells ("no HRS" in E). Bottom: T cells expressing the inhibitory receptor gene near ligand-positive HRS cells ("lig$^+$ HRS" in E) compared to ligand-negative HRS cells ("lig- HRS" in E). Log2 fold changes of values in E are shown. The bar plot is based on all T cells in all regions of an individual patient. The dots indicate the log2 fold change in single regions.

expression of *LAG3* on CD8 T cells and the expression of *CD69* on CD3 T cells, although at a lower frequency, which is in line with the scRNA-seq results (**Fig. 4C**).

CD30 and *LGALS1* were co-expressed more often than background, although this enrichment was not as high as the enrichment of the inhibitory receptors on the T cells (**Fig. 5C**). HLA-DRA expression was the same on *CD30*-expressing cells as background. This means that HRS cells do express *LGALS1* more than other cells, but not *HLA-DRA*. Of note, large variability was observed between patients and regions between the expression of the T and HRS markers with no consistent correlation between the two (**Fig. S6, S7**). This suggests that HRS cells do not consistently induce expression of the inhibitory receptors on T cells across patients and tissue regions.

To see if there was an indication of cell-cell interactions, we analyzed the local enrichment of T cells around HRS cells. We found *CD30*$^+$ cells closely surrounded by *CD69*$^+$ and *LAG3*$^+$ cells (**Fig. 5D**). Therefore, we assessed whether T cells near HRS cells were more or less likely than other T cells to express *CD69* or *LAG3*. In 41 out of 45 regions, T cells expressing *CD69* or *LAG3* were enriched near HRS cells. This indicates that this subset of T cells is either recruited towards HRS, or the expression of the inhibitory receptor is induced near HRS cells. Still, the variation across different regions of the tumors was high in most patients (**Fig. 5E,F**). Then, the enrichment of *CD69*/*LAG3*-expressing T cells was compared between HRS cells that expressed the corresponding ligand and the other HRS cells (**Fig. 5F**). Interestingly, in patient PB26217 T cells near *LGALS1*$^+$ HRS cells expressed *CD69* more often in all tumor regions, but this was more variable for patients PB27302 and PB11473 and not the case for patient PB06422. The expression of *LAG3* on T cells was not enriched near *HLA-DRA*$^+$ HRS cells compared to other HRS cells in any patient.

Discussion

Here, we present a unique method to simultaneously capture scRNA-seq data of HRS and TME cells. While previous cHL scRNA-seq studies only captured TME cells[20,21] and compared their data to publicly available bulk HRS cell profiles of another cohort[19], we could identify constitutively expressed HRS cell membrane protein genes and identify the strength of HRS-immune cell interactions per patient. This approach gave us a possible explanation for the previously observed depletion of $PD1^+$ T cells in cHL tissue[44], namely the depletion of germinal center cell types in cHL tumors including $PD1^+$ TFH cells.

Most interactions were found between HRS and T cells. We found that in pediatric cHL, NK cells and each T cell subset express a different inhibitory receptor gene. In addition, we were able to identify CD69/Galectin-1 and LAG3/HLA-II as a probable interaction between the HRS and T cells in most tumors based on scRNA-seq. Imaging analysis validated the expression of *LAG3* and *CD69* on CD8 and CD3 T cells respectively and indicated that their expression was enriched on T cells that surrounded HRS cells in all patients, although the amount of enrichment varied greatly. However, when also studying the corresponding ligand, the CD69/Galectin-1 interaction was present in one out of four investigated tumors, and the interaction was only identified in a subset of the tissue regions of the other tumors. The LAG3/HLA-II interaction was not observed in any patient. Possibly, other ligands on HRS cells are important for the interaction in the patient without enrichment, or the protein level of the ligands in HRS cells is different from transcript levels. The LAG3/HLA-II and CD69/Galectin-1 interaction might thus not be universally targetable but could pose a potential targetable interaction in a subset of patients. Experiments should validate the *in vivo* protein-protein binding of this ligand-receptor pair and should assess the effect of interfering with this interaction.

By capturing single HRS cells, new potential universal membrane markers could be identified. In addition, in one patient a continuum of the HRS-core transcriptional program could be identified. Extended cohorts capturing more HRS cells should validate this transcriptional heterogeneity and investigate its link with treatment response and prognosis, as this would for example reduce their usefulness as universal HRS markers in flow cytometry. A combination of single-cell DNA and RNA sequencing might elucidate whether the transcriptional heterogeneity can be explained by heterogeneity on the DNA level, like previously indicated for CNVs[47], or whether it is absent as was suggested for patient PB16107.

In addition, most interactions identified in this study were only found in one or a few patients. This highlights the importance of considering the inter- and intra-patient heterogeneity of the cHL TME when investigating new targets for immunotherapies. Finally, some differences were found with previous studies, e.g., a higher fraction of CD8 T cells expressing *LAG3* compared to CD4 T cells, while the opposite was

previously reported in adults[19], calling for a study investigating the differences between childhood/young adult cHL and adult cHL.

Acknowledgements

This work was funded by an ERC consolidator grant from the European Research Council (ERC; no. 864499) to R. van Boxtel and supported by a collaborative grant initiative between the Princess Máxima Center, Utrecht University, and the University Medical Centre Utrecht. Additionally, this work was supported by the Oncode Institute, funding J.K. de Kanter, A. Steemers, N. Groenen, M. Verheul, R. van Boxtel, and F. Meyer-Wentrup. We thank Single Cell Discoveries for their help with library preparation. We thank the Princess Máxima Center Single Cell facility for aiding in the single cell RNA sequencing and performing data processing. Imaging services were performed in The Princess Máxima Imaging Center (Utrecht, the Netherlands). FACS was performed with the assistance of the Princess Máxima Center FACS facility. Data and material were provided by the Biobank Data Access Committee and the Princess Máxima Center genomics core. We thank all patients and parents for donating the tissue.

Author contributions

F.M., F.H., T.M. initiated the project. J.K., A.M.B., T.M., F.M., N.G., R.B. were responsible for the experimental design. J.K., A.S., R.B. drafted the manuscript. A. M.B., N.G., T.M. performed the experimental work for the scRNA-seq. A.S. and R.I. performed the RNAscope and imaging. J.K. performed all data analysis. M.S., L.W., F.M. provided clinical information. F.H., A.R., A.M.B., A.B., T.M., R.B, F.M. supervised the project.

Methods

Patient material

All lymph nodes that were used for scRNA-seq were obtained as frozen single-cell suspensions from the biobank of the Princess Máxima Center for Pediatric Oncology, Utrecht, the Netherlands in accordance with the declaration of Helsinki. The use of the material was approved by the Biobank and Data Access Committee under proposals PMCCRC2018016 (Hodgkin Lymphoma lymph nodes) and under PMCLAB2021-254 (reactive lymph nodes).

Patients were selected that were diagnosed between 2019 and 2022 with cHL of any subtype and for who frozen single cell suspensions of lymph node material was available. After the initial two patients were processed by scRNA-seq (PB24752, PB262127), the other samples were screened for having cells with HRS cell-like marker expression (see below). Only those samples that had those cells were selected.

Immunohistochemistry

Immunohistochemistry (IHC) information was obtained from routine diagnostic IHC. IHC was performed on the Leila Bond III staining system. IHC stainings were analyzed by an experienced pathologist. EBV status of HRS cells was determined by EBER in situ hybridization imaged on the same machine.

Single cell suspension

Single cell suspensions were made by the diagnostic technicians specialized in flow cytometry as follows. Wash medium (20% FCS, 80% RPMI-Glutamax) with 2% gentamycine was added to the lymph node biopsies, which were minced and pushed through a 100um cell strainer, spun down for 10 minutes at 300 RCF at room temperature and resuspended in washing medium and put on ice. Cells were divided over different ampuls, spun down at 469 RCF for 5 minutes, resuspended in 500 ul washing medium, 500ul freezing medium was added in drops (80% washing medium, 20% DMSO) and cells were stored in liquid nitrogen freezers.

Fluorescence-activated cell sorting

Samples were thawed and stained for FACS after which events were sorted in 384 well plates. Sorting was performed on a Sony SH800S Cell Sorter. In all samples, all sorted events were $DAPI^-$negative singlets, as determined by an FSC-H/FSC-A and an SSC-H/SSC-A gate. The Sony SH800S measures backward scatter, not side scatter, but as these are indicative for the same granularity/complexity "SSC" is used for clarity as the abbreviation throughout the manuscript. In addition, for all samples unbiased live singlets were sorted into the majority of the wells. In addition, for all samples, part of the wells was filled with SSC^+ cells. Except for the first two processed cHL lymph node samples, the other seven were stained with fluorescently labeled antibodies against CD20, CD30, CD40, CD95, and CD15. Depending on the number of cells present after thawing and the fractions of cells that were part of these subsets, part of the wells were filled with SSC^+CD20^- cells ("tumor-lenient") and $SSC^+CD20^-CD30^+CD40^+CD95^+CD15^+$ cells ("tumor-strict"). The gating strategy for the HRS cell gate is depicted in **Fig. S1**. Reactive lymph nodes were stained with CD20 and part of the wells were filled with SSC^+ and $CD20^+$ cells. The BioLegend antibodies used were as follows. CD20-BV421 (clone 2H7, 302329, 1:50), CD15-FITC (clone HI98, 301903, 1:50), CD95-PE (clone DX2, 305607, 1:50), CD30-APC (BY88, 333909, 1:25), CD40-AF700 (clone 5C3, 334327, 1:50), CD3-APC/Fire750 (clone SK7, 344839, 1:50). In addition, samples were stained with DAPI (Sigma-Aldrich, D9542-1MG, 500mM 1:250). Reactive lymph nodes were stained with CD20-FITC (clone 2H7, 302303) instead of CD20-BV421, DAPI, and DRAQ5 (50uM 1:100). For a full overview of samples, cell numbers and sorting strategy, see **Table S1**.

Single-cell RNA sequencing library, processing, and filtering

Single-cell RNA sequencing was performed according to the SORT-seq protocol[23]. 384 well plates were filled with Sigma mineral oil (10ul), RT primers (50nl), and External RNA Controls Consortium (ERCC) spike-in transcripts. The first

column was left empty to be able to control for background contamination after sequencing. Library preparation was done as previously described[23,48]. Paired-end 75bp sequencing was performed on an Illumina Nextseq 500. Mapping to reference genome hg38, annotation using Gencode 26, and gene-level transcript quantification was done with the Sharq pipeline[49].

Only wells in which the library was successfully constructed and sequenced, as judged from the ERCC transcripts, were considered. Then DecontX was run, using all the successful wells from all plates, to remove ambient RNA[50]. Finally, wells with less than 1000 transcripts, less than 200 measured genes, or with more than 50% mitochondrial reads were removed. Further processing was done using the Seurat R package v4.1.1[51]. This included normalization to 10,000 transcripts, data scaling, and identification of the 2000 most variable features. Variable features were filtered for cell cycle genes, sex genes, shock protein genes, and ribosomal genes as described before[52]. This resulted in a list of 1798 variable genes. Principal component analysis (PCA, 100 PCs) was performed, and the first 30 principal components were used for UMAP dimensionality reduction and shared nearest neighbor clustering (resolution 0.05).

To identify cell types in the scRNA-seq data, first the HRS cells were identified. Then, the expression of HRS markers per cluster was compared to the HRS expression pattern based on the immunohistochemistry of the pathology department. Then SingleR package v 1.10.0[27] was applied with the celldex v.1.6.0 Monaco reference data. In addition, CHETAH v 1.13.0[26] was run with the default tumor immune reference. Based on these classifiers and canonical marker expression, cell types were assigned to each cluster. Subsequently, the clusters containing T cells were processed separately by Seurat as described above to better define the T cell subtypes. These subtypes were determined by T cell marker expression. The same procedure was performed for all myeloid cells.

Differential composition analysis was performed using the R package DCATS v0.99.6[41] with default settings to determine which cell types were depleted and enriched in cHL compared to RLN. The p-values from the likelihood ratio test were fdr corrected.

Copy number variation

The inferCNV package[28] v 1.12.0 was used to infer CNVs from the scRNA-seq data, using the standard settings "cutoff=0.1, denoise=TRUE, cluster_by_groups=TRUE".

Cell-cell interactions

The immune cell composition of the cHL lymph node from PB24752, in which no HRS cells could be detected, had a high fraction of GC-B and TFH cells, but almost no exhausted T cells. As this makes it likely that the part of the lymph node tissue that

was analyzed had low or no HRS cells and was thus not representative of the tumor tissue, this sample was excluded in all subsequent immune cell analyses.

Receptor-ligand pairs were taken from the curated iCellNet interaction database[53]. Only those interactions were selected for which all ligand and receptor genes were measured in the scRNA-seq data. First, the expression of each ligand and receptor was averaged per cell type per patient. For each patient, only cell types with 5 or more cells were used. Interaction scores were determined by multiplying the averaged ligand expression of one cell type with the averaged receptor expression of another cell type from the same patient. An interaction was considered to be active in a patient when the interaction score was 0.1 or higher. "Common" interaction between HRS and a specific cell type were those that were present in at least all but 2 patients (with a minimal of 3).

Bulk RNA-seq

Bulk RNA sequencing data generated for routine diagnostics were obtained from the Princess Máxima Center biobank under proposals PMCLAB2021-205 and PMCLAB2021-254.
Cell type deconvolution was performed with CIBERSORTx[42]. A signature matrix was constructed from the scRNA-seq data using default settings. Cell type fractions were imputed in "absolute mode" using "S-mode" batch correction with 100 permutations. Differential cell abundance was calculated with the Wilcoxon-test and fdr corrected.

HRS markers

Differential expression analysis (DEA) was performed for the scRNA-seq data using the FindMarkers function from the Seurat package, comparing HRS cells to all other cell types, using the setting "logfc.threshold = 0, min.pct = 0, min.diff.pct = 0, min.cells.feature = 0, min.cells.group = 0". Affymetrix data from Tiacci et al.[13] and Steidl et al.[30] were normalized using RMA (oligo package v1.60.0[54]) and DEA was performed using limma v3.52.2 with standard settings[55]. HRS cells from Steidl et al. were compared to bulk cHL, GC-B cells, and centroblasts. In the data from Tiacci et al. HRS were compared to naive B cells, memory B cells, centrocytes, and centroblasts. In the bulk-RNA seq data, DEA was performed using the DESeq2 v1.36.0 package using standard settings and using data of reactive lymph node samples as the control[56].

For each of the four expression datasets, a gene was considered differentially expressed when the adjusted p-value was lower than 0.01 and the log2 fold change (log2FC) was higher than 0 for the bulk RNA seq, or the average log2FC in the scRNA-seq was higher than 0, or the minimal log2FC of all comparisons with the normal B cell references was greater than 0 for Affymetrix data. HRS core genes were those that were identified in the two Affymetrix datasets and the single-cell RNA seq data.

Bulk RNA sequencing data was also obtained from non-Hodgkin lymphoma samples from the Princess Máxima Center biobank. Deseq2 was applied to perform differential expression analysis between Hodgkin and non-Hodgkin lymphomas as described above.

GO term enrichment was performed using the enrichGO function from ClusterProfiler v.4.4.4. using the "biological process" ontology[57]. KEGG pathway enrichment was performed with the diffEnrich v0.1.2 package with the following setting, "N = 5" [58].

SurfaceGenie[59] was used to extract genes that express a protein that has a high likelihood of being present on the surface of the cell membrane. High likelihood was defined as being predicted as a membrane protein by at least 4 out of 5 methods. Then, HPAanalyze[60] v1.14.0 was used to select genes that were expressed in fewer than 10 out of 127 normal cell types from 55 tissues from the Human Protein Atlas.

WGS, processing, SNV calling, and copy number variation

For patient PB16107, whole genome sequencing was performed on 3,500 bulk-sorted HRS cells. The sorting protocol was the same as described above for the scRNA-seq with cells sorted based on FSC/SSC characteristics and the following staining profile $DAPI^{-}SSC^{+}CD20^{-}CD30^{+}CD40^{+}CD95^{+}CD15^{+}$. DNA was isolated using the NEBNext Ultra II FS DNA Library Prep kit. As a control, bulk T cells ($CD20^{-}CD3^{+}$) were sorted (~500.000 cells) and DNA was isolated with the Qiagen QIAamp DNA micro kit. DNA of HRS and T cells was sent for 30X WGS. FREEC was used to determine the copy number variation in this sample using a bin size of 2Mb[61]. The IAP pipeline was used for read alignment and variant calling and further filtering was performed using SMuRF v2.1.2 as described previously[62]. For each SNV position in the WGS, all scRNA-seq reads from PB16107 that spanned the mutation sites were extracted. SNVs that had at least 80 reads spanning it were selected. From these reads, per cell, the number of alternative and reference reads was determined based on unique UMIs. Finally, per cell the total number of UMIs that supported any of the alternative alleles was calculated.

RNAscope In Situ Hybridization

In situ hybridization assays were performed with RNAscope technology using the RNAscope Fluorescent Multiplex kit v2 (ACD, 323100) and 4-plex Ancillary Kit (ACD, 323120). Formalin-fixed, paraffin-embedded (FFPE) tissues from four cHL patients were cut into 6μm sections using a microtome. Probes used included the following: Hs-TNFRSF8-C1 (ACD, 593451-C1), Hs-LGALS1-C2 (ACD, 486281-C2), Hs-HLA-DRA-C2 (ACD, 475891-C2), Hs-CD69-C3 (ACD, 494471-C3), Hs-CD8A-C3 (ACD, 560391-C3), Hs-LAG3-C4 (ACD, 553931-C4), and Hs-CD3E-C4 (ACD, 553971-C4). FFPE sections were deparaffinized in xylene and rehydrated in ethanol. RNAscope Hydrogen Peroxide was applied to block endogenous peroxidase

activity before target retrieval was performed for 15 min in a preheated glass beaker (100°C) containing target retrieval solution. Protein digestion was then carried out by applying RNAscope Protease Plus. Probes were hybridized for 2 hours at 40°C followed by signal amplification. After amplification, probes were fluorescently labelled with Opal dyes: Opal 520 (Akoya Biosciences, FP1487001KT, 1:1500) was assigned to *HLA-DRA* and *CD3E*, Opal 570 (Akoya Biosciences, FP1488001KT, 1:1500) was assigned to *LGALS1* and *CD8A*, Opal 620 (Akoya Biosciences, FP1495001KT, 1:1500) was assigned to *CD69* and *LAG3*, and Opal 690 (Akoya Biosciences, FP1497001KT, 1:1500) was assigned to *TNFRSF8*. Finally, slides were incubated for 30 sec in DAPI and then coverslipped. Following staining, imaging was performed on a Leica STELLARIS 8 Confocal Microscope with a white light laser (tunable range 440-790 nm) using a 20x/0.75 NA multi-immersion objective set to oil. Tilescan images were acquired with 10% overlap at pixel size 0.142 x 0.142 μm in 16Bit. The tiles were merged in Leica LASX software.

RNAscope data processing

Upper and lower limits were set for each fluorescent label for each slide based on a manual inspection in ImarisViewer. ImarisViewer was also used to select 5-7 regions of approximately equal sizes that encompassed all variability in the slides. In python v3.9.16, the regions were isolated from each image and values outside of the determined limits were capped, and the minimal values were restored to 0. Then, scikit-image (v0.19.2) was used to determine a threshold between true and false positive signals using Otsu's method. Values below this threshold were reduced to 0 to further reduce noise. Then, scikit-image was used to erode and dilute 1px of the remaining values in order to remove the last noise. Finally, the signal of each block of 51px (7.242 μm) was averaged. Subsequent analysis was done in R v4.2.1 and tidyverse 1.3.1. Mean intensities per block were converted to z-scores as z = (intensityblock – min(intensityregion))/sd(intensityregion). Then z-scores were normalized to 0-1. When investigating the z-scores, a bimodal distribution was observed with a minor peak near 0 and a major peak at 0.4. The dip between these peaks was found at 0.14 and values below 0.14 (3.3% of total positive values) were filtered out.

The expression of T cells expressing inhibitory markers was done by dividing the 7.242 μm blocks in three categories, 1) those directly adjacent to blocks with CD30 expression and the interaction ligand (LGALS1/HLA-DRA), 2) those directly adjacent to blocks with CD30 expression, but not the expression of the ligand, 3) those no adjacent to a block with CD30 expression. For of the three categories, in each region in each tissue, the number of blocks with T cells was determined (based on CD3E/CD8A expression). Then, the fraction of these T cell blocks that co-expressed the inhibitory receptor (CD69/LAG3) was determined. To calculation enrichment, the fraction of T cells that expressed the inhibitory receptor was compared between category 1/2 and category 3 (general enrichment around HRS cells) or between category 1 and 2 (enrichment around HRS cells that expressed the interaction ligand).

References

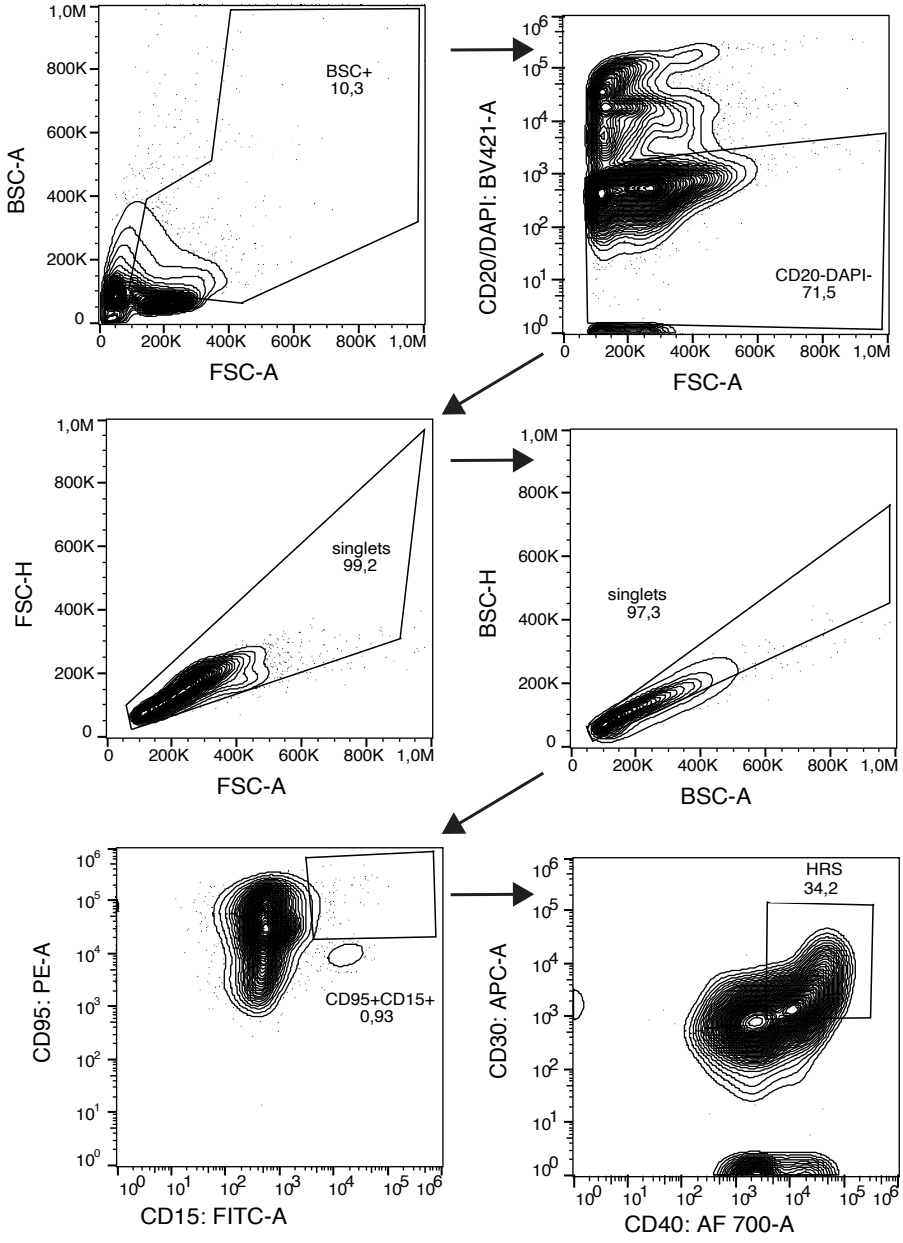
1. Brice, P., de Kerviler, E. & Friedberg, J. W. Classical Hodgkin lymphoma. The Lancet 398, 1518–1527 (2021).
2. LaCasce, A. S. Treating Hodgkin lymphoma in the new millennium: Relapsed and refractory disease. Hematol Oncol 37, 87–91 (2019).
3. Raut, M., Singh, G., Hiscock, I., Sharma, S. & Pilkhwal, N. A systematic literature review of the epidemiology, quality of life, and economic burden, including disease pathways and treatment patterns of relapsed/refractory classical Hodgkin lymphoma. Expert Rev Hematol 15, 607–617 (2022).
4. Ehrhardt, M. J. et al. Integration of Pediatric Hodgkin Lymphoma Treatment and Late Effects Guidelines: Seeing the Forest Beyond the Trees. Journal of the National Comprehensive Cancer Network 19, 755–764 (2021).
5. Ansell, S. M. Hodgkin lymphoma: 2018 update on diagnosis, risk-stratification, and management. Am J Hematol 93, 704–715 (2018).
6. Daw, S. et al. Risk and Response Adapted Treatment Guidelines for Managing First Relapsed and Refractory Classical Hodgkin Lymphoma in Children and Young People. Recommendations from the EuroNet Pediatric Hodgkin Lymphoma Group. Hemasphere 4, e329 (2020).
7. Dixon, S. B. et al. Specific causes of excess late mortality and association with modifiable risk factors among survivors of childhood cancer: a report from the Childhood Cancer Survivor Study cohort. The Lancet 401, 1447–1457 (2023).
8. van Bladel, D. A. G. et al. Novel Approaches in Molecular Characterization of Classical Hodgkin Lymphoma. Cancers (Basel) 14, (2022).
9. Hertel, C. B., Zhou, X., Hamilton-Dutoit, S. J. & Junker, S. Loss of B cell identity correlates with loss of B cell-specific transcription factors in Hodgkin/Reed-Sternberg cells of classical Hodgkin lymphoma. Oncogene 21, 4908–4920 (2002).
10. Maura, F. et al. Molecular Evolution of Classic Hodgkin Lymphoma Revealed Through Whole-Genome Sequencing of Hodgkin and Reed Sternberg Cells. Blood Cancer Discov 4, 208–227 (2023).
11. Marafioti, T. et al. Hodgkin and Reed-Sternberg cells represent an expansion of a single clone originating from a germinal center B-cell with functional immunoglobulin gene rearrangements but defective immunoglobulin transcription. Blood 95, 1443–1450 (2000).
12. Wein, F. & Küppers, R. The role of T cells in the microenvironment of Hodgkin lymphoma. J Leukoc Biol 99, 45–50 (2016).
13. Tiacci, E. et al. Analyzing primary Hodgkin and Reed-Sternberg cells to capture the molecular and cellular pathogenesis of classical Hodgkin lymphoma. Blood 120, 4609–4620 (2012).
14. Kapp, U. et al. Hodgkin's Lymphoma-Derived Tissue Serially Transplanted Into Severe Combined Immunodeficient Mice. Blood 82, 1247–1256 (1993).
15. Nakhoda, S., Rizwan, F., Vistarop, A. & Nejati, R. Updates in the Role of Checkpoint Inhibitor Immunotherapy in Classical Hodgkin’s Lymphoma. Cancers (Basel) 14, (2022).
16. Herrera, A. F. et al. Brentuximab vedotin plus nivolumab after autologous haematopoietic stem-cell transplantation for adult patients with high-risk classic Hodgkin lymphoma: a multicentre, phase 2 trial. Lancet Haematol 10, e14–e23 (2023).
17. Randall, M. P. & Spinner, M. A. Optimizing Treatment for Relapsed/Refractory Classic Hodgkin Lymphoma in the Era of Immunotherapy. Cancers (Basel) 15, (2023).
18. Aoki, T. & Steidl, C. Novel insights into Hodgkin lymphoma biology by single-cell analysis. Blood 141, 1791–1801 (2023).
19. Aoki, T. et al. Single-Cell Transcriptome Analysis Reveals Disease-Defining T-cell Subsets in the Tumor Microenvironment of Classic Hodgkin Lymphoma. Cancer Discov 10, 406–421 (2020).
20. Aoki, T. et al. Single-cell profiling reveals the importance of CXCL13/CXCR5 axis biology in lymphocyte-rich classic Hodgkin lymphoma. Proceedings of the National Academy of Sciences 118, e2105822118 (2021).
21. Veldman, J. et al. CD4+ T cells in classical Hodgkin lymphoma express exhaustion associated transcription factors TOX and TOX2. Oncoimmunology 11, 2033433 (2022).
22. Stewart, B. J. et al. Spatial and molecular profiling of the mononuclear phagocyte network in Classic Hodgkin lymphoma. Blood 141, 2343–2358 (2023).
23. Muraro, M. J. et al. A Single-Cell Transcriptome Atlas of the Human Pancreas. Cell Syst 3, 385–394 (2016).
24. Fromm, J. R. & Wood, B. L. A Six-Color Flow Cytometry Assay for Immunophenotyping Classical Hodgkin Lymphoma in Lymph Nodes. Am J Clin Pathol 141, 388–396 (2014).
25. Fromm, J. R. & Wood, B. L. Strategies for immunophenotyping and purifying classical Hodgkin lymphoma cells from lymph nodes by flow cytometry and flow cytometric cell sorting. Methods 57, 368–375 (2012).
26. de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. P. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. Nucleic Acids Res 47, e95–e95 (2019).

27. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 20, 163–172 (2019).
28. Tickle, T., Tirosh, I., Georgescu, C., Brown, M. & Haas, B. inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV.
29. Wienand, K. et al. Genomic analyses of flow-sorted Hodgkin Reed-Sternberg cells reveal complementary mechanisms of immune evasion. Blood Adv 3, 4065–4080 (2019).
30. Steidl, C. et al. Gene expression profiling of microdissected Hodgkin Reed-Sternberg cells correlates with treatment outcome in classical Hodgkin lymphoma. Blood 120, 3530–3540 (2012).
31. Weniger, M. A. & Küppers, R. Molecular biology of Hodgkin lymphoma. Leukemia 35, 968–981 (2021).
32. Gholiha, A. R. et al. Revisiting IL-6 expression in the tumor microenvironment of classical Hodgkin lymphoma. Blood Adv 5, 1671–1681 (2021).
33. Niens, M. et al. TARC and MDC Are the Only Chemokines with Highly Increased Levels in Serum of Hodgkin Lymphoma Patients. Blood 108, 2268 (2006).
34. Von Hoff, L. et al. Autocrine LTA signaling drives NF-κB and JAK-STAT activity and myeloid gene expression in Hodgkin lymphoma. Blood 133, 1489–1494 (2019).
35. Kapp, U. et al. Interleukin 13 Is Secreted by and Stimulates the Growth of Hodgkin and Reed-Sternberg Cells. Journal of Experimental Medicine 189, 1939–1946 (1999).
36. Newcom, S. R., Ansari, A. A. & Gu, L. Interleukin-4 Is an Autocrine Growth Factor Secreted by the L-428 Reed-Sternberg Cell. Blood 79, 191–197 (1992).
37. Skinnider, B. F. & Mak, T. W. The role of cytokines in classical Hodgkin lymphoma. Blood 99, 4283–4297 (2002).
38. Vinatzer, U. et al. Mucosa-Associated Lymphoid Tissue Lymphoma: Novel Translocations Including Rearrangements of ODZ2, JMJD2C, and CNN3. Clinical Cancer Research 14, 6426–6431 (2008).
39. Ansell, S. M. et al. Overall Survival with Brentuximab Vedotin in Stage III or IV Hodgkin's Lymphoma. New England Journal of Medicine 387, 310–320 (2022).
40. Castellino, S. M. et al. Brentuximab Vedotin with Chemotherapy in Pediatric High-Risk Hodgkin's Lymphoma. New England Journal of Medicine 387, 1649–1660 (2022).
41. Lin, X., Chau, C., Ma, K., Huang, Y. & Ho, J. W. K. DCATS: differential composition analysis for flexible single-cell experimental designs. Genome Biol 24, 151 (2023).
42. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 12, 453–457 (2015).
43. Green, M. R. et al. Integrative analysis reveals selective 9p24.1 amplification, increased PD-1 ligand expression, and further induction via JAK2 in nodular sclerosing Hodgkin lymphoma and primary mediastinal large B-cell lymphoma. Blood 116, 3268–3277 (2010).
44. Patel, S. S. et al. The microenvironmental niche in classic Hodgkin lymphoma is enriched for CTLA-4–positive T cells that are PD-1–negative. Blood 134, 2059–2069 (2019).
45. de la Fuente, H. et al. The Leukocyte Activation Receptor CD69 Controls T Cell Differentiation through Its Interaction with Galectin-1. Mol Cell Biol 34, 2479–2487 (2014).
46. Koyama-Nasu, R. et al. The cellular and molecular basis of CD69 function in anti-tumor immunity. Int Immunol 34, 555–561 (2022).
47. Mangano, C. et al. Precise detection of genomic imbalances at single-cell resolution reveals intra-patient heterogeneity in Hodgkin's lymphoma. Blood Cancer J 9, (2019).
48. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol 17, 77 (2016).
49. Candelli, T. et al. Sharq, a versatile preprocessing and QC pipeline for Single Cell RNA-seq. bioRxiv (2018) doi:10.1101/250811.
50. Yang, S. et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol 21, 57 (2020).
51. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 33, 495–502 (2015).
52. Calandrini, C. et al. An organoid biobank for childhood kidney cancers that captures disease and tissue heterogeneity. Nat Commun 11, 1310 (2020).
53. Noël, F. et al. Dissection of intercellular communication using the transcriptome-based framework ICELLNET. Nat Commun 12, 1089 (2021).
54. Carvalho, B. S. & Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. Bioinformatics 26, 2363–2367 (2010).
55. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43, e47–e47 (2015).
56. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data

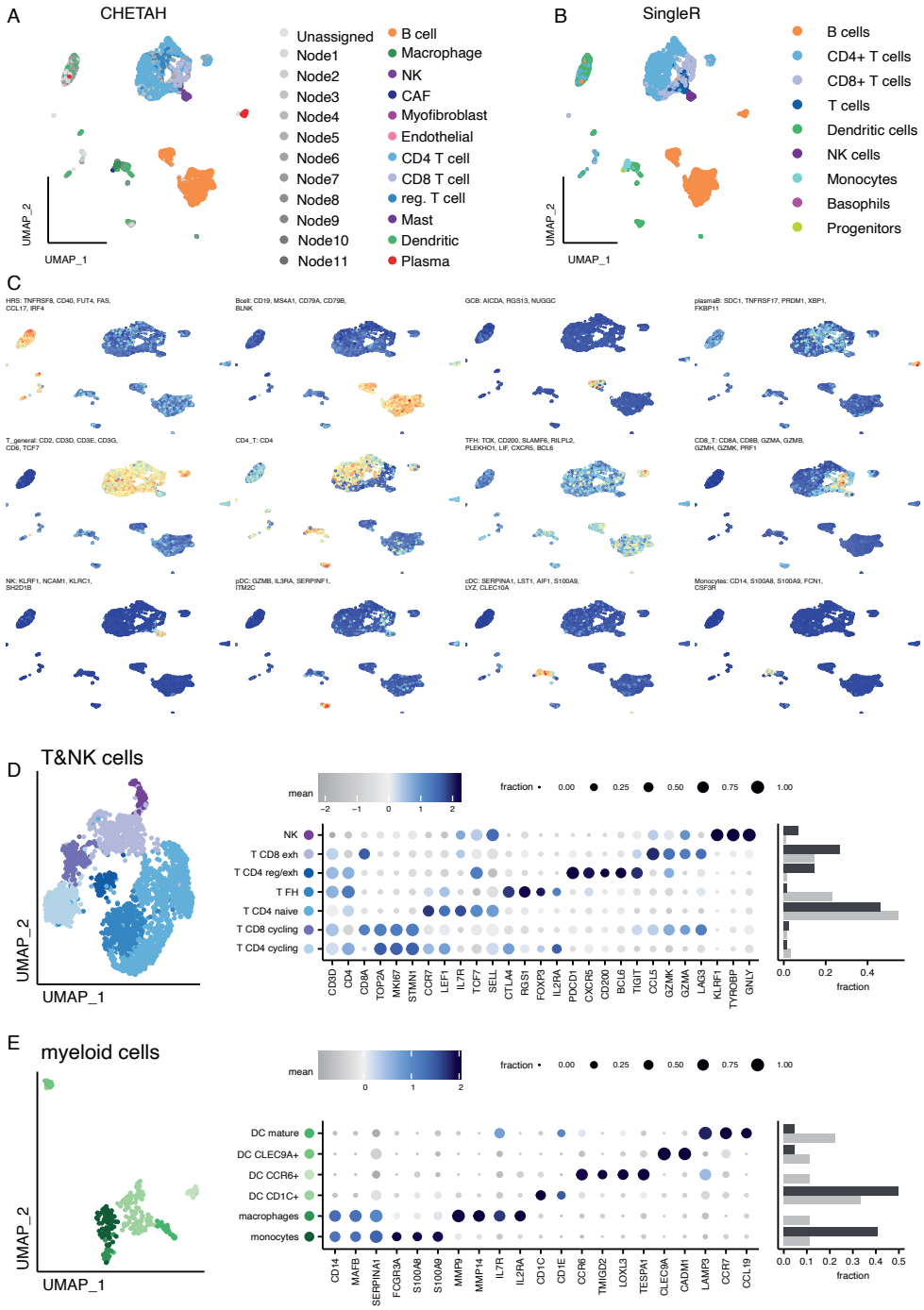
with DESeq2. Genome Biol 15, 550 (2014).
57. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2, (2021).
58. Petersen, D. R. et al. Elevated Nrf-2 responses are insufficient to mitigate protein carbonylation in hepatospecific PTEN deletion mice. PLoS One 13, e0198139- (2018).
59. Waas, M. et al. SurfaceGenie: a web-based application for prioritizing cell-type-specific marker candidates. Bioinformatics 36, 3447–3456 (2020).
60. Tran, A. N., Dussaq, A. M., Kennell, T., Willey, C. D. & Hjelmeland, A. B. HPAanalyze: an R package that facilitates the retrieval and analysis of the Human Protein Atlas data. BMC Bioinformatics 20, 463 (2019).
61. Boeva, V. et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics 28, 423–425 (2012).
62. Bertrums, E. J. M. et al. Elevated Mutational Age in Blood of Children Treated for Cancer Contributes to TherapyRelated Myeloid Neoplasms. Cancer Discov 12, 1860–1872 (2022).

Supplementary Material

A



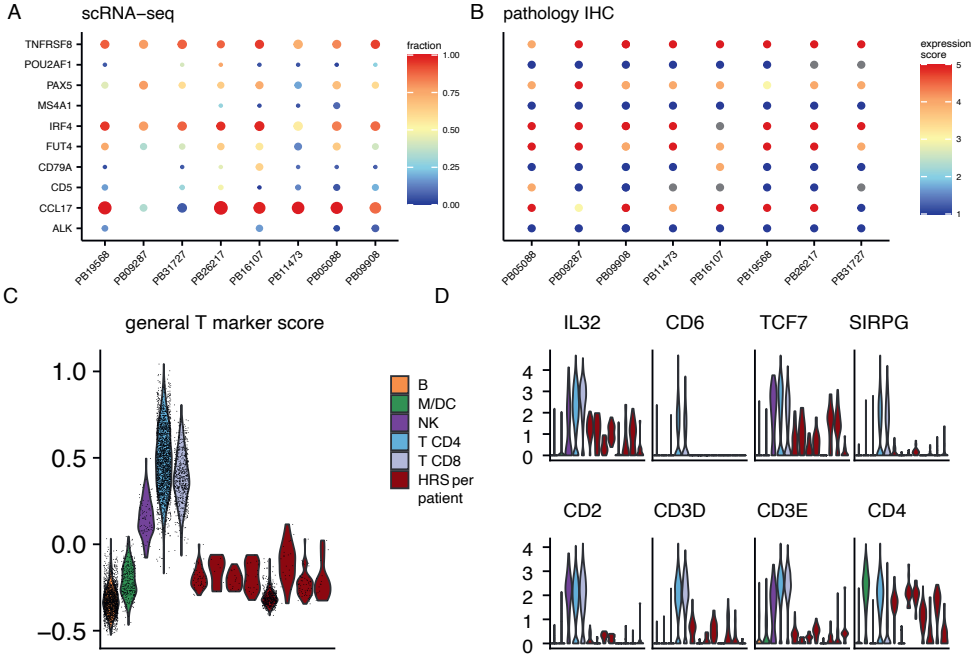
Supplementary Figure 1. Flow cytometry sorting strategy
A) A representative plot for the sorting strategy for HRS cells.



Supplementary Figure 2. The identification of immune cell types in the scRNA-seq data
A) UMAP plot with cells labeled with cell types labels generated by the CHETAH algorithm using the default reference data. Node[x] labels indicate intermediate assignments in the hierarchical classification

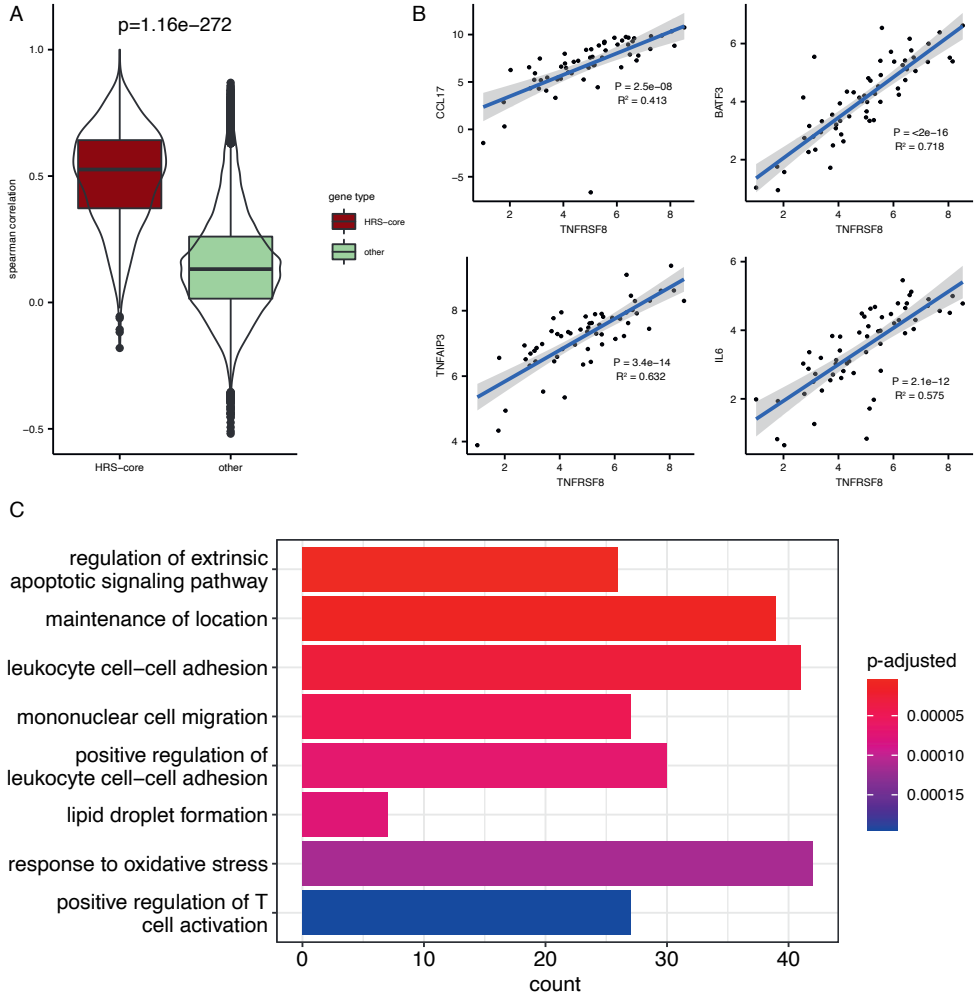
The legend continues on the next page.

model. **B)** UMAP plot labeled with cell type labels generated using SingleR with the Monaco et al. reference from celldex. **C)** UMAP plots colored by different Seurat Module scores (normalized mean scores) of canonical immune cell markers. **D)** The expression of markers in the NK and T cell subset of the scRNA-seq data. A UMAP plot of the NK and T cells, processed separately (left). A dotplot showing the fraction of each cell types expressing genes and the mean expression in these cells (middle). A comparison of the percentage of each cell type in cHL lymph nodes compared to reactive lymph nodes. **E)** Similar to (D), but for myeloid cells and marker genes.

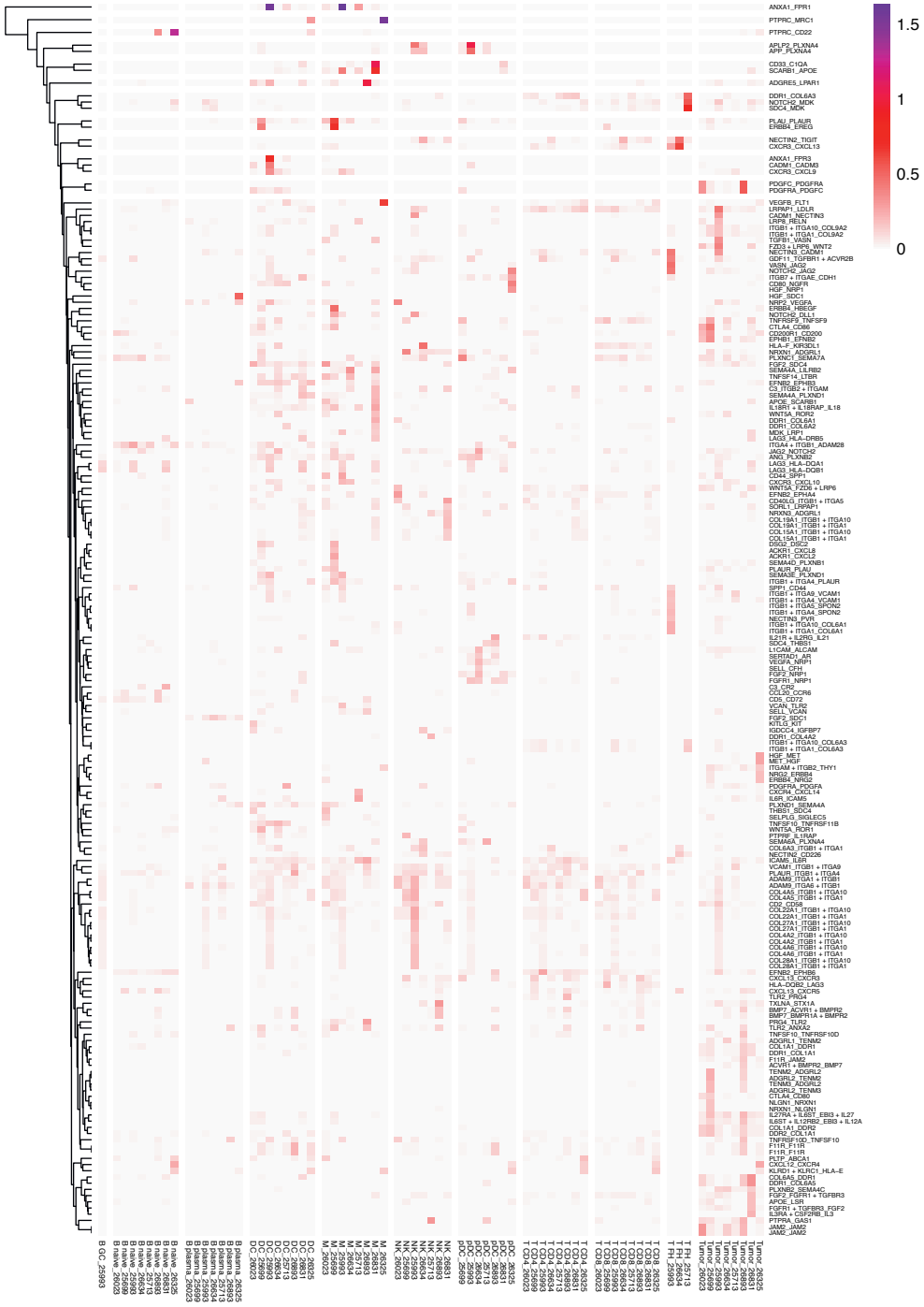


Supplementary Figure 3. scRNA-seq HRS cell expression is in line with IHC-defined HRS markers

A) A dotplot of the scRNA-seq expression of markers used by pathologists to distinguish lymphoma subtypes. **B)** A dotplot of protein expression as determined by routine diagnostics immunohistochemistry and quantified by a visual scoring from a experienced pathologist. Scores between 1 (not present) and 5 (highly expressed) were given. **C)** A violin plot of the Seurat Module score (normalized average) of genes differentially expressed by T cells in this data, clustered by cell type. The HRS cells of each patient were plotted separately. **D)** Violin plots of T cell marker genes, clustered by cell type. The HRS cells of each patient were plotted separately.

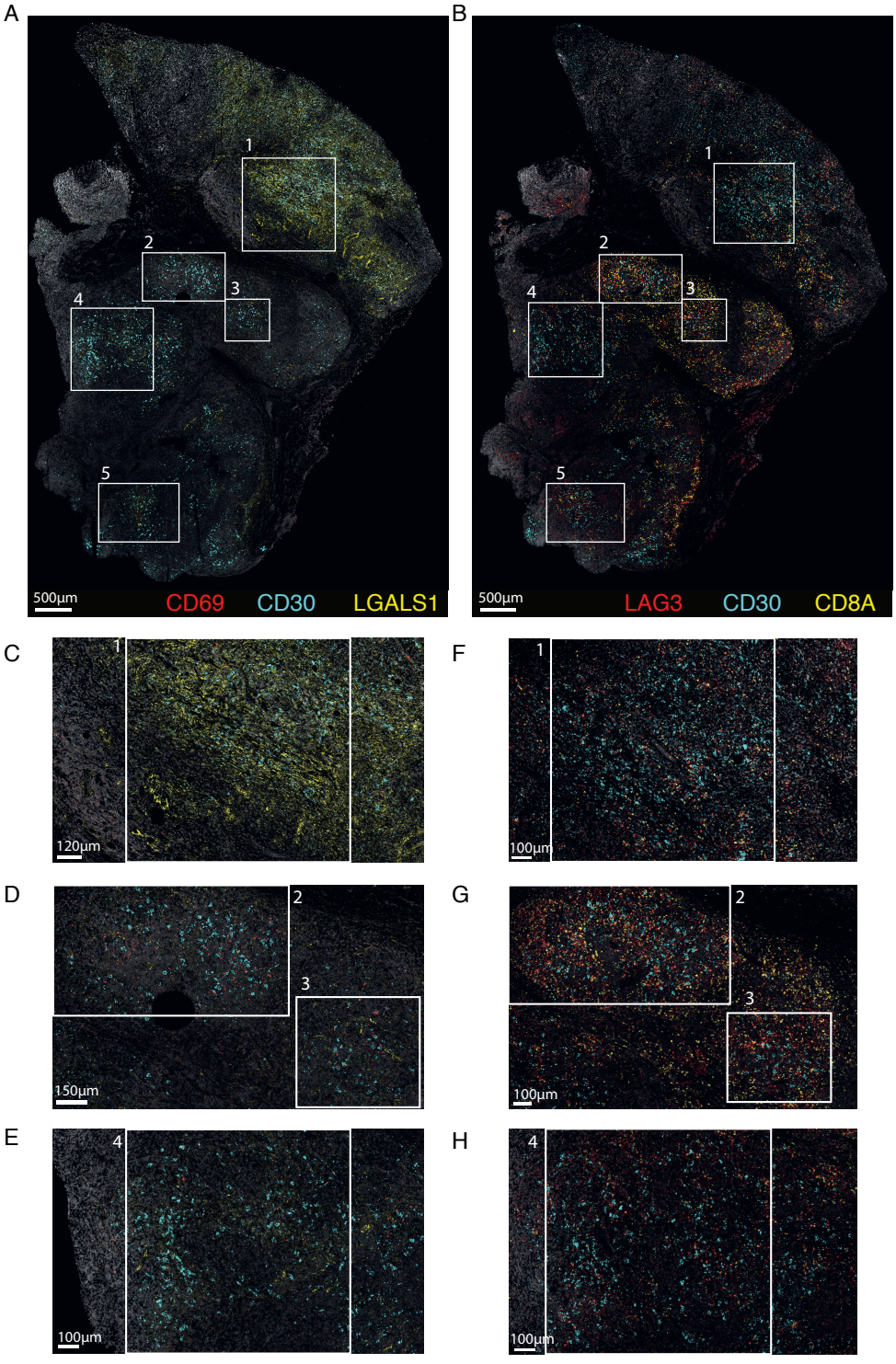


Supplementary Figure 4. HRS-core gene characteristics.
A) Boxplots of the spearman correlation between the expression of genes and the expression of TNFRSF8 (CD30) in bulk RNA-seq data of cHL lymph nodes. HRS-core genes and other genes are plotted separately. **B)** Examples of the correlation between HRS-core genes and TNFRSF8 (CD30) in bulk RNA-seq of cHL lymph nodes. Each dot is a sample. **C)** A bar plot of GO-terms enriched in the HRS-core genes.



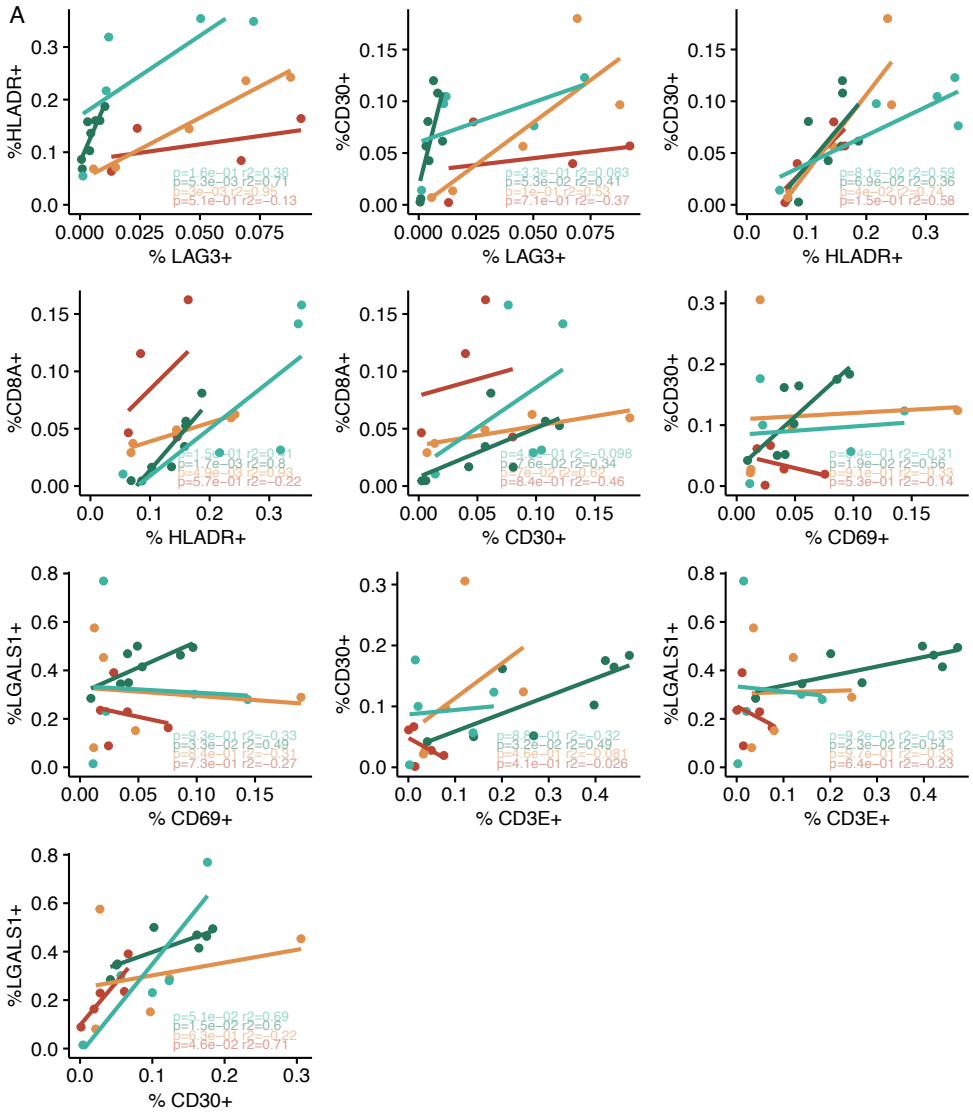
Supplementary Figure 5. Most HRS-immune cell interactions are only found in one or a few patients.

A) Similar to Figure 4A, but interactions present in less than 3 samples are shown.


A
1
2
3
4
5
500µm
CD69
CD30
LGALS1
B
1
2
3
4
5
500µm
LAG3
CD30
CD8A
C
1
120µm
D
2
3
150µm
E
4
100µm
F
1
100µm
G
2
3
100µm
H
4
100µm

Supplementary Figure 6. T and HRS marker gene expression does not correlate across tumor regions

A) The entire RNAscope image of panel 1 of patient PB06422. The regions used for analyses are indicated in white circles. Zoom-in images of these regions are shown in C-E. **B)** The entire RNAscope image of panel 2 of patient PB06422. The regions used for analyses are indicated in white circles. Zoom-in images of these regions are shown in F-H. **C-E)** Zoomed in images of the regions indicated in A. **F-H)** Zoomed in images of regions indicated in B.



Supplementary Figure 7. A lack of correlation between T and HRS marker expression over regions

A) Similar to Figure 5A, but for all correlation not shown there.

Supplementary Table 1. Number and type of events that were sorted, sequenced, and passed QC per sample.

patient	Wells sorted	Wells pass	CD20	Live singlet	SSC+	SSC+CD20-	Tumor strict
PB19568	738	480	0	204	67	114	95
PB09287	738	389	0	225	51	72	41
PB31727	738	336	0	167	58	74	37
PB26217	738	483	0	440	43	0	0
PB16107	1107	858	0	322	65	0	471
PB11473	738	396	0	217	44	94	41
PB05088	738	324	0	132	42	55	95
PB09908	738	441	0	188	57	80	116
PB24752	1107	499	0	710	28	0	0
PB25394	738	382	71	226	85	0	0
PB32331	738	384	67	226	91	0	0
PB32684	738	499	84	312	103	0	0


8

General discussion

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

Since the introduction of chemotherapy, survival rates of pediatric cancer patients have greatly improved[1–3]. However, the pace of improvement has steadily declined for multiple cancer types, e.g., acute lymphoblastic leukemia (ALL)[4,5]. In addition, the burden of late effects in pediatric cancer survivors remains high[6,7]. Currently, these two problems are addressed through the two following approaches. First, the dose of some chemotherapeutic drugs is reduced in subgroups of patients who have good survival rates, even after dose reduction[6,8,9]. For this, currently administered compounds are identified that are toxic to normal tissues and thus candidates for dose reduction. Second, therapies are developed that specifically target malignant cells. These have reduced toxicity in normal tissues compared to conventional chemotherapy and should thus cause fewer late effects. In addition, these targeted therapies have the potential to further improve survival rates, primarily in high-risk patient groups[10,11]. Although efforts in these two areas have been ongoing for multiple decades, chemotherapy usage as a cornerstone for cancer therapy has not declined[6] and the number of patients that are eligible for treatment with targeted therapies remains low[12,13]. The work described in this thesis aims to expand our molecular understanding of late effects and therapy targets. In the long term, this could contribute to a reduction of late effects in pediatric cancer survivors. To this end, recently developed single-cell transcriptomics and genomics approaches were applied. In this chapter, the results of the work described in this thesis are combined and discussed, follow-up studies are proposed, the effectiveness of the work is evaluated, and the potential fundamental and clinical impact of the work is reviewed.

Which drugs mutate which cells? Looking beyond a single tissue or cohort.

In **chapter 2**, we directly answered the important questions of whether, and to what extent, chemotherapy can mutate the DNA of normal cells in vivo. We showed that thiopurines and platinum compounds are directly mutagenic to healthy cells. In addition, we found two mutational signatures (SBSB and SBSC) for which the causative treatment still needs to be identified. Adding additional cases in **chapter 3** revealed an additional signature, SBSG. This signature was present in three patients who had a primary Ewing sarcoma and SBSG is therefore likely linked to Ewing sarcoma treatment. A genetic study from the United States of pediatric, therapy-related myeloid neoplasms (t-MN) also identified the first two mutagenic drug groups but did not identify signatures SBSB, SBSC, or SBSG[14]. Another recent study, of adult t-MN, on the other hand, found an additional mutational signature linked to melphalan treatment, a drug not administered to the patients described in **chapter 2**[15]. In conclusion, the administered drugs, and therefore identifiable signatures, differ per cohort. To determine the in vivo mutagenicity of every drug, more studies similar to the ones mentioned above should be conducted. These should focus on studying cohorts with varying primary cancer types, patient ages, and geographical locations, as treatment protocols vary over all these groups. In addition, **chapter 4** describes that ganciclovir, an antiviral nucleoside analog, is mutagenetic to healthy cells, highlighting that such cohort studies should be extended beyond the context of cancer.

The data generated in chapters 2, 3, and 4 were all derived from a single tissue, i.e., blood. Are the results therefore only applicable to this tissue, or can they be generalized to the rest of the body? Chapter 2 revealed that platinum compounds are mutagenic to all exposed normal and leukemic blood cells. Previous research has indicated that these drugs are also mutagenic in normal liver cells, in normal colon cells[16], and in cancer metastases of a large variety of tissue types[17]. In this specific case, the results from **chapter 2** can therefore be extrapolated to most or all other tissues. In contrast, while in vitro exposure to 5-FU induces signature SBS17 mutations, this footprint is not consistently found in cancer and normal cells after exposure to the drug in vivo[16,18–20]. This observation indicates that 5-FU, and perhaps also other drugs, might be mutagenic to only a subset of cell types, cell states, or tissues. The source of the studied cells can also be important, for example in the context of hematopoietic stem cell transplantation (HSCT). Here, only the donated blood cells undergo the transplantation procedure, while only the recipient's tissues are exposed to the conditioning regimen. In **chapter 4**, we showed that the HSCT procedure is not associated with mutagenesis in the transplanted blood cells and is therefore genetically safe for the donated blood. In contrast, in a therapy-related myeloid neoplasm (t-MN) patient described in **chapter 2**, the recipient's hematopoietic stem and progenitor cells (HSPCs) accumulated additional mutations after each unsuccessful round of allogeneic HSCT. These mutations were of a unique signature "SBSB" and were therefore likely a consequence of the conditioning regimen. So, while we concluded in **chapter 4** that the procedure is safe for the donated blood cells, all the recipient's cells can still be mutated by the procedure, which may contribute to late effects. Taken together, findings of mutagenicity studies performed in a single tissue and setting are not necessarily directly translatable to other tissues/settings. It is therefore important that studies into the in vivo mutagenicity of therapeutic compounds, as proposed in the previous paragraph, are performed in a variety of tissues. Ideally, a large study would be conducted that, post-mortem, collects many normal tissues of patients treated with a large variety of treatments and compares treatment-induced mutations.

How chemotherapies drive t-MN: selection, mutational aging, and drivers

In chapters 2 to 4 we studied the mechanisms by which exposure of healthy cells to chemotherapies and nucleoside analogs can contribute to the development of t-MN. One of the ways this can occur is by altering selective pressures. In **chapter 3**, we described how the end of platinum compound exposure is the rate-limiting step for the development of TP53-proficient pediatric t-MN. These t-MN only started to expand after the cessation of platinum drug administration. The expansion of TP53-deficient cells, as observed in Li-Fraumeni patients who carry a TP53 germline mutation, was less inhibited by platinum compounds.

Besides changing selective pressures, exposure to chemotherapeutic drugs also induces mutations in normal cells. Mutations accumulate in healthy cells during a

person's lifetime, a phenomenon known as mutational aging[21]. The extent to which mutational aging contributes to biological aging is not fully understood. Neither is it known whether this might be accelerated by chemotherapy-induced mutations. An undisputed process by which mutations contribute to aging-associated diseases is by causing cancer-driver mutations. Chemotherapy accelerates this process. For example, chemotherapy induces single nucleotide cancer driver mutations in t-MN (**chapter 2**). More importantly, many of the fusion genes that are the main driver of t-MN are likely chemotherapy-induced (**chapter 2**). Interestingly, the number of translocations that were found in normal cells in these patients was low. This would suggest that chemotherapy-induced translocations are a rare event and thus the rate-limiting step in t-MN development. However, it is possible that most normal cells that acquire a chemotherapy-induced structural variant either do not survive in vivo or are not able to clonally expand in vitro. In both cases, they would not be captured in our studies. PTA is mainly applied to study t-MN blasts in **chapter 3**. However, it could be applied to normal cells as well, which would remove the in vitro selection bias in future studies.

Whether chemotherapy-induced passenger mutations, which don't drive cancer, also contribute to biological aging is less clear. In **chapter 2**, the number of mutations detected in the normal HSPCs of a few pediatric cancer patients was in the range of healthy elderly individuals. Despite childhood cancer survivors experiencing an earlier onset of chronic health conditions, their individual physiological aging after treatment does not entirely mirror that of an elderly person. Interestingly, carriers of germline aberrations in the proofreading domain of POLE, which encodes for DNA polymerase ε, can even accumulate up to ten times the normal mutation load in their cells, to levels never observed during normal aging. Still, they have normal rates of biological aging, other than an increased cancer risk[21]. Therefore, there seems to be no correlation between the absolute number of mutations in cells and biological aging in these two patient groups. There are multiple possible explanations for this lack of a correlation.

First, the type of mutations that accumulate might be important. In human tissues, mutational signatures SBS1, SBS5, and SBS18 accumulate at a constant rate in tissues over a person's lifespan and are therefore called "clock-like"[22–24]. These signatures are also observed during aging in different mammalian species[25]. The longer the life span of a species, the slower the accumulation of these signatures is. This suggests that these types of mutations are somehow connected to biological aging. The mutations caused by chemotherapy generally do not have the same mutational profile as clock-like signatures and might therefore have a different effect on cells. The same holds true for mutations in POLE germline mutation carriers.

Second, the effect of accumulated mutations on aging might not be direct and immediate. Recent studies have shown that during aging, tissues like the skin

and esophagus are taken over by clones that harbor cancer-driving mutations[26,27]. These clones can have an altered or decreased functioning, leading to age-related diseases[28–30]. The additional chemotherapy-induced mutations in childhood cancer survivors might increase the number of cells that harbor cancer-driving mutations. These could outcompete neighboring cells over time resulting in more and larger clones at an earlier age. However, this process takes time. The latency between the initial mutation and the clonal outgrowth and resulting disease might therefore explain why not all effects of chemotherapy are immediate. This can relatively easily be investigated by performing the clonal analysis used in the abovementioned studies on tissues of childhood cancer survivors and comparing the number and size of clones with data from healthy adults of similar ages.

Finally, chemotherapy-induced DNA damage might induce biological aging via other mechanisms than mutations. For clarity, DNA damage (or a DNA lesion) is a physical abnormality in the DNA, like a DNA break, an abnormal base like 8-oxoguanine, or a DNA adduct. A somatic mutation on the other hand is a DNA sequence in which nucleotide(s) are changed compared to the germline, which can be the consequence of an unrepaired or incorrectly repaired DNA lesion that mismatches during replication. The presence of DNA damage in a cell can trigger pathways that induce senescence and inflammation, both of which are linked to aging[31]. In addition, the activity of the DNA damage repair (DDR) machinery can have other consequences besides DNA repair. For example, the DDR machinery can permanently alter the epigenetic markers surrounding the original damage, which has been linked to aging[31]. In conclusion, not only mutations, but also epigenetic changes, senescence, and inflammation are caused by DNA damage and can contribute to aging.

Remarkably, it is estimated that the DNA in most cells of the body is damaged 10^5 times per day under normal circumstances[31]. Still, healthy cell accumulates 10-60 clock-like mutations per year per cell. The estimated error rate of the DDR machinery is therefore ~10^{-6} [24]. It seems likely that the fraction of drug-induced DNA lesions that is not (correctly) repaired and thus results in a mutation is higher compared to aging-related DNA lesions. Ganciclovir (GCV) is antiviral nucleoside analog drug that induces single base substitutions. Lesions segregation analysis in **chapter 4** indicated that tens to hundreds of GCV DNA lesions remained unrepaired and resulted in mutations during one HSPC cell cycle. This means that only in the unlikely case that GCV caused 10^7-10^8 DNA lesions during this time, would the DNA repair efficiency of GCV be the same as that of aging-related DNA lesions. Equally high numbers of segregating lesions were observed after in vitro exposure of iPSCs to multiple exogenous compounds among which chemotherapeutic drugs[32]. The potentially lower efficiency of the DDR machinery to repair exogenous DNA damage could be because the DDR is optimized during evolution for repairing aging-associated DNA lesions. Alternatively, the DDR could already be saturated by the repair of aging-associated damage, and any additional damage is repaired less efficiently.

If the DDR machinery indeed has a lower efficiency of repairing drug-induced compared to aging-induced DNA lesions, this would be relevant for the link between the number of drug-induced mutations in a cell and biological aging. Under this hypothesis, a cell would have undergone less DNA damage and less DNA repair when it has 100 drug-induced mutations in its genome compared to 100 age-related mutations. Such a cell would therefore have accumulated fewer of the changes induced by DNA damage and the DDR machinery, such as the epigenetic changes, inflammation and senescence described above. In other words, when we identify a child in **chapter 2** that has mutation loads in their HSPCs that are similar to those of an elderly person due to chemotherapy-induced mutations, the amount of DNA damage and thus age-related physiological changes that the cell underwent would still be lower than the cell of an elderly person, and its functioning would therefore be better. This would explain the lack of correlation between the number of mutations and biological aging in these children.

A similar explanation could underly the normal rate of biological aging in carriers of proofreading mutations in POLE, in whom mutations accumulate faster due to polymerase mistakes not being repaired. In these patients, the rate of DNA damage and the level of DDR activity do not differ with normal aging. Therefore, the rate of physiological cell changes, other than DNA mutations, and therefore the rate of biological aging is the same as in healthy individuals. This theory might also be relevant for pediatric HSCT recipients of older (e.g., parental) stem cell donors. Even though the HSCT procedure does not generally induce mutations (**chapter 4**), the transplanted blood of an older donor harbors more age-related mutations, and therefore also the physiological cell alterations associated with aging, than a child's blood. Indeed, in most HSCT studies, younger donor age is associated with better overall and disease-free survival, both in adult and pediatric recipients[33–37].

In conclusion, chemotherapeutics and other drugs can contribute to different steps of t-MN development. How treatments affect t-MN development likely depends on which drugs are administered, at which concentrations, the patient's age, and the presence of germline mutations. Indeed, 24% of children with t-MN carried cancer predisposition mutations, whereas in primary cancers this is 8.5%[38]. Although this work investigated some of these components for specific contexts, such as a single treatment or specific germline mutation, in many other contexts the role of and interplay between these components is still not unraveled. For example, while we showed that exposure to platinum compounds can inhibit t-MN expansion (**chapter 3**), other drugs may have the same effect. Also, we showed an interaction between germline TP53 mutations and platinum compound exposure, but likely more of such genotype-exposure interactions exit. Finally, potential synergistic interactions between classes of chemotherapeutic drugs that promote t-MN development are yet to be found. For example, hypothetically, the combination of topoisomerase inhibitors followed by platinum compounds might be synergistic in inducing t-MN,

as the former can induce fusion genes, and the latter can prevent clonal expansions of cells that harbor such a driver until the many other cells in the microenvironment are killed by the chemotherapy and the cell has a highly competitive advantage.

How to find therapy targets: Hodgkin Lymphoma as an example

Targeted therapies pose an alternative to conventional chemotherapy, and they have the potential to increase survival while decreasing the toxic exposure of normal cells. In **chapter 7**, potential targets of pediatric Hodgkin lymphoma were investigated by single-cell RNA sequencing (scRNA-seq), and it applies the scRNA-seq cell type classifier described in **chapter 6**. Although the newly described CD69-galectin 1 interaction could be a potential therapy target, the study indicated that the interaction was not present in all patients. The interaction could thus not be used to develop a new general therapy. Recently, a WGS study of isolated HRS cells was published that focused on the developmental trajectory of HL[39]. Hopefully, more such studies will be carried out, as they could identify genomic targets that are more consistently affected and are essential for malignant cell survival, like the NF-κB or JAK/STAT pathways that were identified in previous studies[40,41].

How fundamental research could influence clinical care: a long and difficult road

The ultimate long-term goal of the work described in this thesis is to contribute to improved clinical care that results in a reduced chance that cancer survivors develop late adverse effects. Some of the findings presented here could be of consequence to the patient, but the amount of additional evidence needed to move towards the clinic, and the conditions needed for actual change in clinical practice are vast.

As discussed previously, one of the main aims of this work is to identify genotoxic drugs that could be candidates for dose reduction or replacement. One of the most mutagenic groups of drugs identified by us (**chapter 2, 3**) and others is platinum compounds. Platinum compounds have been increasingly administered to cancer patients in the last few decades[6,14,17]. When used, they are essential for the effectiveness of the regimen. Therefore, dose reduction has only been attempted in patient groups that have a very good outcome and that have tumors that respond well to chemo- and radiotherapy such as HPV-associated oropharyngeal cancer[42]. However, even in these patients, full replacement of cisplatin with another drug results in a significantly worse outcome[42]. GCV was another highly mutagenic compound identified in this work (**chapter 4**). GCV and its prodrug valganciclovir are used to treat and prevent CMV disease in immunocompromised individuals, who sometimes receive GCV prophylactically for extensive periods of time[43]. In this setting, the question arises whether the prevention of a potential infection outweighs the additional risk that comes with mutating many cells in the patient's body. Stopping the prophylactic use of GCV will be taken as an example to explore the steps between the fundamental research described in this thesis and changes in clinical practice.

First, it is essential to quantify the additional risks of developing late effects that may be associated with GCV treatment. To reliably determine the relative risk for, e.g., cancer, the presence of GCV mutations must be assessed in larger numbers of patients compared to the few patients in **chapter 4**. As WGS, and certainly single-cell WGS, is an expensive technique to apply to thousands of samples, alternative approaches should be taken. Primarily, the power of large, existing databases could be leveraged. A recent study searched for the mutational signature of GCV in two targeted sequencing databases covering more than 100,000 samples and identified 22 samples with GCV-induced mutations[44]. This study confirmed the overrepresentation of drivers in RAS-family genes in these 22 cancers and gained additional evidence that GCV is involved in inducing these cancer driver mutations. This approach is very promising for assessing the mutagenicity of other drugs in a population-wide manner. However, there are a few limitations to this approach. First, such studies are only possible when a drug induces a specific mutation (e.g. T>G) in a limited number of contexts. For example, GCV induces only C>A mutations and these were only induced in four out of sixteen possible trinucleotide contexts. Similarly, the majority of mutations that are induced by molnupiravir (**chapter 4**), platinum compounds (**chapter 2**), and thiopurines (**chapter 2**) occur in a few specific trinucleotide contexts. Some other drugs like zidovudine (**chapter 4**) induce mutations of all mutation types in all contexts with approximately equal likelihood. Other drugs induce mutations that mimic clock-like signatures (**chapter 2**). In both cases, it is impossible to pinpoint these drugs' exact contribution to carcinogenesis using mutational signature analysis as their mutations are indistinguishable from aging-associated mutations. Second, because sequencing is mainly performed in healthy individuals or in the context of cancer, for most other patient groups there are no large sequencing databases. Because no large sequencing databases exist for patients with viral infections, tracing the effect of antiviral drugs will be more difficult. Third, comprehensive clinical metadata including a registry of all administered drugs is needed to link drugs to mutations and late effects. Even in the study with more than 100,000 samples, no cancer risk estimation could be made for the use of GCV. The reason was that only the cancer treatment, and not the use of other drugs, like GCV, were recorded in the investigated sequencing studies. Therefore, it was impossible to calculate the fraction of GCV-treated patients that developed cancer and compare it to a reference population. Gathering such detailed information on such a scale is highly difficult and will remain a significant obstacle in research into the involvement of drugs like GCV in late effects, even though the number of patients included in (sequencing) databases is growing constantly.

A more conventional step in the research of carcinogenicity is an in vivo rodent study. Both the EMA and FDA report that systemic exposure to GCV can lead to embryo growth retardation in pregnant mice and decreased fertility in males and that GCV is mutagenic and carcinogenic in animal studies in dose ranges comparable to human administration[45,46]. However, as is inherent to mice experiments, how the cancer risk in rodents translates to the human setting remains unknown.

If a significant additional risk for late effects after GCV treatment exists and can be proven, the second step would be to assess the feasibility of discontinuing prophylactic use of GCV. More than half of the people in high-income countries are seropositive for CMV[47], and when CMV disease develops in immunocompromised patients most die of the infection when it is untreated[48]. Prophylactic GCV treatment is successful and reduces CMV disease prevalence to approximately 6% of CMV-infected patients[49,50]. However, preemptive treatment, i.e., treatment only after detection of viral DNA or protein in the blood, is similarly effective in preventing symptomatic CMV infections in transplant recipients compared to general prophylactic treatments for all patients and is now preferred in most centers[49–51]. This approach already decreased the number of patients that receive GCV. Interestingly, preemptive use of the non-mutagenic foscarnet (**chapter 4**) results in similar survival as GCV[52]. However, even though GCV induces neutropenia in one-third of the recipients, the severe metabolic and nephrological toxicity associated with foscarnet is the reason that it is mostly used as a second-line treatment of CMV[52,53]. Of note, the prophylactic use of (val)acyclovir, which was not mutagenic in our screen (**chapter 5**), was described as less effective in early studies but has recently been shown to be as effective as GCV treatment in subgroups of transplant recipients[53,54]. However, it is associated with severe psychiatric side effects. Finally, in 2017 letermovir, an inhibitor of the CMV terminase complex, was approved as a CMV prophylaxis and was shown to be similarly effective as GCV, while inducing less neutropenia[55,56]. However, the cost of letermovir is currently an order of magnitude higher than the cost of GCV[57]. In conclusion, there are potential alternatives to the prophylactic use of GCV, but the decreased risk of long-term side effects needs to outweigh the additional acute side effects or additional costs.

The last step in implementing clinical change based on the results of **chapter 4** would be a long-term follow-up study that compares the number of infections, the survival, the costs, the acute side effects, and the long-term side effects of an alternative approach to preemptive GCV. Such a study would take at least a decade and would be resource-intensive. It is thus important that great care is taken in the assessment of the potential impact, feasibility, and costs at each subsequent step of the research pipeline. This way, the increasing costs of each subsequent study are spent effectively.

How impactful and disruptive is fundamental research: money well spent?

When discussing the effectiveness of the entire clinical research pipeline, it is important to include an evaluation of pre-clinical research, like the work presented here. The effectiveness of fundamental research could be defined as the quality of the newly obtained fundamental knowledge compared to the money and time invested in the research. High-quality fundamental knowledge can achieve two things. Either it is in line with previous work and incrementally adds knowledge and theory, or it disrupts existing theories and renders previous work obsolete. Within the (bio-) medical field, disruptive work has the potential to induce clinical changes although this can take a long time, as has been discussed above. According to a recent study based on paper citation networks, the disruptiveness of research has steadily declined over the past 8 decades[58]. This emphasizes the need for constant monitoring of the (cost-)effectiveness of one's research. As WGS currently costs 1000 euros for 30x coverage of one sample, the sequencing costs of the WGS studies described here approach 50,000 euros, which does not include the costs of wet lab experiments, salaries, and other expenses. The results of these studies might be categorized as introducing incremental change rather than being disruptive. For example, thiopurines and platinum compounds were previously shown to be mutagenic in cancer cells and were expected to also mutate healthy cells (**chapter 2**)[17,18,59]. Similarly, clonal hematopoiesis in adults after treatment with platinum compounds is enriched in TP53 mutations, suggesting that cells with those mutations were selected by this treatment (**chapter 3**)[60]. When purely focusing on the research into childhood cancer treatment with the most impact (either clinical or social), one could argue that the resources for this research would have been more effectively spent in fields that have a more direct, and more substantial impact on the clinic, like the optimization of personalized dosing[61,62].

Still, WGS studies have the potential to contribute to clinical practice by improving diagnosis and prognosis by identifying new drivers and subgroups with different survival rates. In addition, even though the studies in chapters 2 and 3 did not pose a revolutionary or disruptive idea, they confirmed common theories, something that is essential to moving the research field forward. Furthermore, GCV was previously proven to be mutagenic and carcinogenic in mice, but to prove its mutagenicity in un-infected human cells in vivo was important and, although not disruptive in the setting of fundamental science, has the potential to lead to significant clinical impact. If the prophylactic use of GCV would change, this would affect a significant part of the 80,000 yearly recipients of HSCT and 140,000 yearly recipients of solid organ transplants[63,64]. Still, at the moment of writing, the **chapter 4** article has mostly been cited in the context of genomic consequences of HSCT, but a few times in the context of GCV safety. The absence of more follow-up or attention might be a consequence of the results not reaching the target audience of clinical researchers. Alternatively, the absence of solid clinical data in the article and the lack of a risk assessment make the clinical relevance smaller. So, even though the work presented here mostly confirms

previous theories and is therefore not disruptive in nature, the research findings are important and could contribute to clinical impact when sufficiently followed up in future research.

Most importantly, it is critical that researchers (and funding agencies) continuously evaluate the social, clinical, and scientific impact and the novelty of their work and compare it to other lines of research. Inherent to fundamental science is the fact that the outcome is unknown and that sometimes accidental findings drive the biggest changes. For example, **chapter 4** initially focused on the effect of hematopoietic stem cell transplantation on donated blood cells, and the mutagenicity of ganciclovir was an accidental finding. It is partly these kinds of findings that can move research and, eventually, the clinic forward. However, the unpredictability of fundamental research does not mean that scientists cannot reflect on the (maximum) potential fundamental, social, and clinical impact of their work and think about other techniques, approaches, or even other questions that could improve their impact. In this way, scientific progress and effectiveness can be maintained or even improved.

Concluding remarks: limitations and possibilities

Late effects severely decrease the quality of life of pediatric cancer survivors. Improving our understanding of the molecular mechanisms by which chemotherapies lead to late effects could further improve the survivors' quality of life. Genomic studies, like those presented here, can teach us multiple things about second cancers, and how chemotherapies contribute to them. First, they aid in the characterization of the driver alterations and therefore diagnosis, prognosis, and risk stratification. In addition, genomic studies that identify mutational signatures and link them to drugs are creating a constantly growing "map" of mutagenic compounds and their mutational signatures. WGS studies of time-series samples and single-cell WGS are unraveling the timing of second cancer development relative to the primary diagnosis and treatment. This can help to better determine the role of the treatment and to identify the rate-limiting steps in the second cancer development. The next step in these studies is to identify the association between drugs, germline variants, tissue types, and/or patient groups. There are however limitations to what these kinds of genomic studies can teach us about second cancer development. In some t-MN, expansion only happens after the acquisition of a second driver besides a fusion gene. In other t-MN, the gene fusion is the only detectible driver. Was the clonal advantage of this fusion big enough for the clone to expand or was it a change in the epigenetic state of the precursor cell, or a change in its environment in which it had a higher clonal advantage? Are similar steps needed in t-MN with multiple drivers, or are the drivers enough in those cases? Other techniques are needed to answer these types of questions. In addition, even though this knowledge is important to understand the molecular mechanisms that lead to t-MN, the direct clinical impact of the work is limited, as most studied drugs are cornerstone treatments in cancer therapies. Studies, like the one described in **chapter 7**, are thus important to finally develop novel targeted therapies that could substitute chemotherapeutic drugs.

In addition, the techniques described in this paper can also be applied to study how mutation accumulation and clonality can contribute to late effects other than cancer, by repeating the clonality studies from skin, liver, and esophagus in cancer survivors and comparing the outcomes to the published data on people who did not undergo treatment[26,30,65]. The main limitation here is the availability of material. Whereas blood (the main material studied in this thesis) can be obtained relatively noninvasively, this is more difficult, or impossible, for other non-malignant tissues.

In conclusion, exposure to chemotherapeutic drugs has many effects on healthy cells. DNA damage and mutations play a large role, but other mechanisms are likely involved. Also, how the changes in normal cells lead to the development of late effects is complex and many factors influence this development. Recently developed genomic methods help us to significantly improve our understanding of this process. However, there is a limit to what they can teach us about late effects, and they should therefore be combined with other techniques. Constant reflection of the scientific community on the subjects and techniques that will have the highest chance of gaining the most important fundamental knowledge with the largest social and clinical impact is essential to ensure that research funds are spent effectively, and patients are helped as best as possible.

References

1. Youlden, D. R. et al. Childhood cancer survival and avoided deaths in Australia, 1983–2016. Paediatr Perinat Epidemiol 37, 81–91 (2023).
2. Trama, A. et al. Is the cancer survival improvement in European and American adolescent and young adults still lagging behind that in children? Pediatr Blood Cancer 66, (2019).
3. Gatta, G. et al. Childhood cancer survival in Europe 1999-2007: Results of EUROCARE-5-a population-based study. Lancet Oncol 15, 35–47 (2014).
4. Rostgaard, K. et al. Survival after cancer in children, adolescents and young adults in the Nordic countries from 1980 to 2013. Br J Cancer 121, 1079–1084 (2019).
5. Sasaki, K. et al. Acute lymphoblastic leukemia: A population-based study of outcome in the United States based on the surveillance, epidemiology, and end results (SEER) database, 1980–2017. Am J Hematol 96, 650–658 (2021).
6. Turcotte, L. M. et al. Temporal Trends in Treatment and Subsequent Neoplasm Risk Among 5-Year Survivors of Childhood Cancer, 1970-2015. JAMA 317, 814 (2017).
7. Gibson, T. M. et al. Temporal patterns in the risk of chronic health conditions in survivors of childhood cancer diagnosed 1970–99: a report from the Childhood Cancer Survivor Study cohort. Lancet Oncol 19, 1590–1601 (2018).
8. Vora, A. et al. Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. Lancet Oncol 14, 199–209 (2013).
9. Getz, K. D. et al. Cytarabine dose reduction in patients with low-risk acute myeloid leukemia: A report from the Children's Oncology Group. Pediatr Blood Cancer 69, e29313 (2022).
10. Bedard, P. L., Hyman, D. M., Davids, M. S. & Siu, L. L. Small molecules, big impact: 20 years of targeted therapy in oncology. The Lancet 395, 1078–1088 (2020).
11. Thomas, A., Teicher, B. A. & Hassan, R. Antibody–drug conjugates for cancer therapy. Lancet Oncol 17, e254–e262 (2016).
12. Haslam, A., Kim, M. S. & Prasad, V. Updated estimates of eligibility for and response to genome-targeted oncology drugs among US cancer patients, 2006-2020. Annals of Oncology 32, 926–932 (2021).
13. Marquart, J., Chen, E. Y. & Prasad, V. Estimation of the Percentage of US Patients With Cancer Who Benefit From Genome-Driven Oncology. JAMA Oncol 4, 1093–1098 (2018).
14. Schwartz, J. R. et al. The acquisition of molecular drivers in pediatric therapy-related myeloid neoplasms. Nat Commun 12, 985 (2021).
15. Diamond, B. et al. Tracking the evolution of therapy-related myeloid neoplasms using chemotherapy signatures. Blood 141, 2359–2371 (2023).
16. Kuijk, E., Kranenburg, O., Cuppen, E. & Van Hoeck, A. Common anti-cancer therapies induce somatic mutations in stem cells of healthy tissue. Nat Commun 13, 5915 (2022).
17. Pich, O. et al. The mutational footprints of cancer therapies. Nat Genet 51, 1732–1740 (2019).
18. Pich, O. et al. The evolution of hematopoietic cells under cancer therapy. Nat Commun 12, 1–11 (2021).
19. Angus, L. et al. The genomic landscape of metastatic breast cancer highlights changes in mutation and signature frequencies. Nat Genet 51, 1450–1458 (2019).
20. Christensen, S. et al. 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. Nat Commun 10, 4571 (2019).
21. Robinson, P. S. et al. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. Nat Genet 53, 1434–1442 (2021).
22. Jager, M. et al. Measuring mutation accumulation in single human adult stem cells by whole-genome sequencing of organoid cultures. Nat Protoc 13, 59–78 (2018).
23. Osorio, F. G. et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep 25, 2308-2316.e4 (2018).
24. Manders, F., van Boxtel, R. & Middelkamp, S. The Dynamics of Somatic Mutagenesis During Life in Humans. Frontiers in Aging 2, (2021).
25. Cagan, A. et al. Somatic mutation rates scale with lifespan across mammals. Nature 604, 517–524 (2022).
26. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. Science (1979) 362, 911–917 (2018).
27. Fowler, J. C. et al. Selection of Oncogenic Mutant Clones in Normal Human Skin Varies with Body Site. Cancer Discov 11, 340–361 (2021).
28. Ng, S. W. K. et al. Convergent somatic mutations in metabolism genes in chronic liver disease. Nature 598, 473–478 (2021).
29. Jaiswal, S. & Libby, P. Clonal haematopoiesis: connecting ageing and inflammation in cardiovascular disease.

Nat Rev Cardiol 17, 137–144 (2020).
30. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature 574, 538–542 (2019).
31. 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).
32. Aitken, S. J. et al. Pervasive lesion segregation shapes cancer genome evolution. Nature 583, 265–270 (2020).
33. González-Vicent, M. et al. Donor age matters in T-cell depleted haploidentical hematopoietic stem cell transplantation in pediatric patients: Faster immune reconstitution using younger donors. Leuk Res 57, 60–64 (2017).
34. Canaani, J. et al. Donor age determines outcome in acute leukemia patients over 40 undergoing haploidentical hematopoietic cell transplantation. Am J Hematol 93, 246–253 (2018).
35. Bastida, J. M. et al. Influence of donor age in allogeneic stem cell transplant outcome in acute myeloid leukemia and myelodisplastic syndrome. Leuk Res 39, 828–834 (2015).
36. Mehta, J. et al. Does younger donor age affect the outcome of reduced-intensity allogeneic hematopoietic stem cell transplantation for hematologic malignancies beneficially? Bone Marrow Transplant 38, 95–100 (2006).
37. Rezvani, A. R. et al. Impact of Donor Age on Outcome after Allogeneic Hematopoietic Cell Transplantation. Biology of Blood and Marrow Transplantation 21, 105–112 (2015).
38. Zhang, J. et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. New England Journal of Medicine 373, 2336–2346 (2015).
39. Maura, F. et al. Molecular Evolution of Classic Hodgkin Lymphoma Revealed Through Whole-Genome Sequencing of Hodgkin and Reed Sternberg Cells. Blood Cancer Discov 4, 208–227 (2023).
40. Liu, Y. et al. The mutational landscape of Hodgkin lymphoma cell lines determined by whole-exome sequencing. Leukemia 28, 2241–2272 (2014).
41. Reichel, J. et al. Flow sorting and exome sequencing reveal the oncogenome of primary Hodgkin and Reed-Sternberg cells. Blood 125, 1061–1072 (2015).
42. Patel, R. R. et al. De-intensification of therapy in human papillomavirus associated oropharyngeal cancer: A systematic review of prospective trials. Oral Oncol 103, 104608 (2020).
43. Humar, A. et al. The Efficacy and Safety of 200 Days Valganciclovir Cytomegalovirus Prophylaxis in High-Risk Kidney Transplant Recipients. American Journal of Transplantation 10, 1228–1237 (2010).
44. Fang, H. et al. Ganciclovir-induced mutations are present in a diverse spectrum of post-transplant malignancies. Genome Med 14, 124 (2022).
45. FDA injection label. https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/209347lbl.pdf (2017).
46. EMA CHMP review of Cymevene. https://www.ema.europa.eu/en/medicines/human/referrals/cymevene (2016).
47. Beam, E. & Razonable, R. R. Cytomegalovirus in Solid Organ Transplantation: Epidemiology, Prevention, and Treatment. Curr Infect Dis Rep 14, 633–641 (2012).
48. Rondeau, E. et al. Effect of prophylactic ganciclovir on cytomegalovirus infection in renal transplant recipients. Nephrology Dialysis Transplantation 8, 858–862 (1993).
49. Khoury, J. A. et al. Prophylactic Versus Preemptive Oral Valganciclovir for the Management of Cytomegalovirus Infection in Adult Renal Transplant Recipients. American Journal of Transplantation 6, 2134–2143 (2006).
50. Small, L. N., Lau, J. & Snydman, D. R. Preventing Post-Organ Transplantation Cytomegalovirus Disease with Ganciclovir: A Meta-Analysis Comparing Prophylactic and Preemptive Therapies. Clinical Infectious Diseases 43, 869–880 (2006).
51. Bhat, V., Joshi, A., Sarode, R. & Chavan, P. Cytomegalovirus infection in the bone marrow transplant patient. World J Transplant 5, 287 (2015).
52. Bacigalupo, A., Boyd, A., Slipper, J., Curtis, J. & Clissold, S. Foscarnet in the management of cytomegalovirus infections in hematopoietic stem cell transplant patients. Expert Rev Anti Infect Ther 10, 1249–1264 (2012).
53. Kotton, C. N. et al. The Third International Consensus Guidelines on the Management of Cytomegalovirus in Solid-organ Transplantation. Transplantation 102, (2018).
54. Reischig, T. et al. Valacyclovir Prophylaxis Versus Preemptive Valganciclovir Therapy to Prevent Cytomegalovirus Disease After Renal Transplantation. American Journal of Transplantation 8, 69–77 (2008).
55. Derigs, P. et al. Letermovir prophylaxis is effective in preventing cytomegalovirus reactivation after allogeneic hematopoietic cell transplantation: single-center real-world data. Ann Hematol 100, 2087–2093 (2021).
56. Limaye, A. P. et al. Letermovir vs Valganciclovir for Prophylaxis of Cytomegalovirus in High-Risk Kidney Transplant Recipients. JAMA 330, 33 (2023).
57. Borsani, O. et al. Comparing Costs Associated to Letermovir Prophylaxis Vs Ganciclovir Pre-Emptive Therapy in HCMV-Seropositive Patients Who Undergone to Hemopoietic Stem Cells Transplantation. Blood 134, 5643 (2019).

58. Park, M., Leahey, E. & Funk, R. J. Papers and patents are becoming less disruptive over time. Nature 613, 138–144 (2023).
59. Li, B. et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. Blood 135, 41–55 (2020).
60. Bolton, K. L. et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nat Genet 52, 1219–1226 (2020).
61. Al-Metwali, B. & Mulla, H. Personalised dosing of medicines for children. Journal of Pharmacy and Pharmacology 69, 514–524 (2017).
62. Tucker, G. T. Personalized Drug Dosage – Closing the Loop. Pharm Res 34, 1539–1543 (2017).
63. 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 (2022).
64. Global Observatory on Donation and Transplantation (GODT). https://www.transplant-observatory.org/. https://www.transplant-observatory.org/ (2023).
65. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science (1979) 348, 880–886 (2015).

Nederlandse samenvatting

DNA-schade veroorzaakt kanker

Kanker is de belangrijkste doodsoorzaak in Nederland. Kanker ontstaat door schade in het DNA van gezonde cellen. DNA is een code met instructies voor het maken van eiwitten. Elke cel bevat een kopie van hetzelfde DNA, een code van 3 miljard moleculen, "nucleotide" genaamd. De hele DNA-code samen wordt het "genoom" genoemd en ieder stuk DNA dat voor een eiwit codeert heet een "gen". Als cellen delen, moet het genoom gekopieerd worden. Dit kopiëren verloopt niet foutloos, waardoor veranderingen in het DNA ontstaan. Naast kopieerfouten kunnen DNA-veranderingen ook ontstaan door stoffen of stralingen die het DNA beschadigen, zoals sigarettenrook en UV licht. Veranderingen in het DNA heten "mutaties". Het grootste deel van de mutaties wordt gerepareerd in de cel, maar ook deze reparatie gebeurt niet altijd correct. Hierdoor krijgt elke gezonde cel in ons lichaam elk jaar enkele tientallen tot honderden mutaties erbij. De meeste mutaties hebben geen groot effect op hoe een cel functioneert. Alleen mutaties in het DNA dat voor eiwitten codeert, kunnen veranderen hoe een cel functioneert. Kanker ontstaat als een cel één of meerdere mutaties krijgt die zorgen dat de cel ongecontroleerd gaat delen. Deze mutaties worden "drivers" of oncogene mutaties genoemd.

De overlevingskans en behandeling van kinderkanker

Doordat kankercellen snel delen, kunnen ze gezonde cellen in de weg zitten waardoor organen steeds minder goed kunnen functioneren, waardoor iemand uiteindelijk kan overlijden. De behandeling van kanker is de afgelopen zeven decennia effectiever geworden. Dit is zeker het geval bij kinderen met kanker. Waar zeventig jaar geleden slechts 10% van de kinderen met kanker overleefden, is dat nu meer dan 80%. Chemotherapie, radiotherapie (bestraling) en chirurgie zijn de meest voorkomende behandelingen van kanker. Kinderen kunnen een hogere dosis chemotherapie verdragen. Hierdoor kunnen kankercellen beter gedood worden. Onder andere daardoor genezen kinderen vaker van kanker dan volwassenen.

Chemotherapie veroorzaakt late effecten

De hoge dosis chemotherapie, die aan kinderen gegeven wordt, heeft een keerzijde. Chemotherapie en ook radiotherapie beschadigen het DNA van cellen. Dit gebeurt niet alleen in kankercellen, maar ook in gezonde cellen. DNA-schade in gezonde cellen kan zorgen voor acute bijwerkingen, zoals haarverlies door het doodgaan van cellen in de haarzakjes. Ook komen er veel bijwerkingen op de lange termijn voor. Zulke bijwerkingen worden "late effecten" genoemd en kunnen jaren tot decennia na de kankerbehandeling optreden en/of aanhouden. Onder late effecten vallen onder andere onvruchtbaarheid en hart- en nierschade. Ook hebben overlevenden van kanker grotere kans op het krijgen van een nieuwe kanker dan gezonde mensen. Doordat kinderen met kanker eerder in hun leven radio- en chemotherapie krijgen dan volwassen patiënten, hebben late effecten langere tijd om te ontwikkelen. Dit,

in combinatie met de hogere dosis en het feit dat de kinderen in ontwikkeling zijn tijdens de behandeling, zorgt ervoor dat late effecten het meeste voorkomen bij patiënten die kinderkanker overleefd hebben.

Het verminderen van late effecten

De hoeveelheid late effecten na kankerbehandeling is in de laatste decennia verlaagd door radiotherapie specifieker op de tumor te richten waardoor de schade aan gezonde cellen beperkt werd. Ook is het in sommige kleine groepen patiënten met een goede prognose mogelijk geweest om de dosis van alle chemotherapie te verlagen zonder dat het de overlevingskans verslechterde. Bij de meeste kinderen met kanker is dit echter niet mogelijk. Daarom zijn andere manieren nodig om late effecten te verminderen.

Een van de manieren om late effecten te verminderen is door de dosis te verlagen van enkel die chemotherapiemedicijnen die het meest schadelijk zijn voor gezonde cellen. Daarvoor moet eerst bekend zijn welke medicijnen het meest schadelijk zijn. De studies die in hoofdstuk 2 tot en met 5 van dit proefschrift beschreven worden, onderzoeken welke medicijnen en behandelingen de meeste schade in het DNA van gezonde cellen veroorzaken. Hierbij wordt gebruikt gemaakt van het feit dat ieder proces dat DNA-schade veroorzaakt, dat doet in een specifiek patroon in het DNA. Dit is te vergelijken met voetafdrukken. Elk dier heeft een andere poot, voet, of hoef, en laat doordoor een andere voetafdruk achter in de grond als het loopt. Mutagenen "voetafdrukken" kunnen gebruikt worden voor het bepalen van de oorzaak van de DNA-schade die in een cel voorkomt. Om deze mutaties te kunnen detecteren worden verschillende recent ontwikkelde technieken toegepast die het mogelijk maken van het gehele genoom (alle 3 miljard nucleotiden) van een enkele cel uit te lezen, "sequencen" genoemd.

Platina en thiopurine medicijnen veroorzaken tweede kankers door DNA-schade

Hoofdstuk 2 van dit proefschrift beschrijft onderzoek dat bestudeert welke chemotherapie medicijnen de meeste DNA-schade in gezonde bloedcellen kan veroorzaken en daarbij kan leiden tot tweede kankers in het bloed. In bijna alle patiënten die behandeld waren met chemotherapie was de hoeveelheid mutaties in het DNA van bloedcellen hoger dan in niet behandelde personen. Sommige cellen hadden een verhoogde hoeveelheid mutaties, maar hadden mutatiepatronen die leken op gezonde cellen. Wat het exacte mechanisme is dat de extra mutaties veroorzaakt in deze cellen is nog onbekend. In de andere cellen bleken voornamelijk twee groepen medicijnen mutaties te hebben veroorzaakt, "thiopurines" en platina bevattende medicijnen. Deze medicijnen bleken ook mutaties veroorzaakt te hebben die leidden tot het vormen van een nieuwe kanker in het bloed. In de toekomst zou dus gekeken kunnen worden of specifiek van deze medicijnen de dosis verlaagd kan worden voor het verminderen van late effecten.

In hoofdstuk 3 wordt een vervolgonderzoek op hoofdstuk 2 beschreven dat aantoont dat platina bevattende medicijnen, naast het vormen van oncogene mutaties, nog een tweede rol hebben in het vormen van tweede kankers. Bij de meeste patiënten bleek dat een beschadigde bloedcel niet snel kon gaan delen zolang platina bevattende medicijnen gegeven werd. Pas wanneer de behandeling met platina medicijnen stopte, ging de beschadigde cel snel delen en werd het een tweede kanker. Er was echter één patiënt waarbij de tweede kanker tijdens de toediening van het platina bevattende medicijn snel ging delen en kanker werd. Deze patiënt bleek een mutatie in het TP53 gen te bevatten in alle cellen in het lichaam, een zogenaamde "kiembaanmutatie". Met celexperimenten wordt in hoofdstuk 3 aangetoond dat cellen met TP53 mutaties inderdaad tijdens behandeling met een platina bevattend medicijn kunnen delen. Voor patiënten met een TP53 kiembaanmutatie zou het daarom belangrijk kunnen zijn om al vroeg tijdens de behandeling te testen voor tweede kankers.

Antivirale medicijnen kunnen het DNA van gezonde cellen beschadigen

Bij sommige agressieve vormen van kanker in het bloed is de enige effectieve behandeling het doodmaken van alle bloedcellen van een patiënt, en het vervangen met het bloed van een donor. Dit heet een stamceltransplantatie. Hoofdstuk 4 toont aan dat deze behandeling geen extra mutaties veroorzaakt in de gedoneerde cellen, iets wat daarvoor onbekend was. Wel bleek door dit onderzoek dat er extra mutaties te vinden waren in de bloedcellen van patiënten die een virale infectie hadden gekregen na de stamceltransplantatie. Deze patiënten waren behandeld met het medicijn ganciclovir, een antiviraal middel dat bij de categorie "nucleotide analogen" hoort. Ganciclovir bleek grote hoeveelheden mutaties veroorzaakt te hebben in de gedoneerde, gezonde cellen van deze patiënten. Tijdens het bestuderen van grote genetische datasets van kankers werden ook ganciclovir-geïnduceerde mutaties gevonden. Sommige van deze mutaties hadden bijgedragen aan het ontstaan van de kanker. De behandeling met ganciclovir, in ieder geval in enkele gevallen, kan dus ernstige late effecten veroorzaken. Hoofdstuk 5 beschrijft een onderzoek waarin de DNA-schade in gezonde cellen wordt onderzocht na blootstelling aan veertien andere "nucleotide analogen" medicijnen. Uit deze studie blijkt dat waarschijnlijk zes van deze veertien medicijnen ook mutaties kunnen veroorzaken, maar allemaal veel minder dan ganciclovir. Het zou daarom de prioriteit moeten hebben om te onderzoeken aan welke patiënten minder ganciclovir gegeven kan worden.

Op zoek naar doelgerichte therapie voor Hodgkin Lymfoom

Voor de behandeling van kinderen met Hodgkin Lymfoom (HL), een vorm van kanker in het bloed, zijn grote hoeveelheden radio- en chemotherapie nodig. Deze kinderen krijgen een van de hoogste aantallen late effecten van alle kinderkankerpatiënten. Daarom is het belangrijk om medicijnen voor HL te identificeren die specifieker de tumorcellen doden en minder schadelijk zijn voor gezonde cellen. Zulke therapieën worden "doelgerichte therapieën" genoemd. Om "doelen" voor deze medicijnen te vinden kan de activiteit van alle genen ("expressie") per cel uitgelezen worden,

om zo genen met een verhoogde activiteit in kankercellen te vinden. Deze techniek resulteert in een zogenaamd genexpressie profiel van duizenden cellen. De eerste stap van de analyse van deze data is het bepalen van het celtype waarvan elk profiel afkomstig is. Hoofdstuk 7 beschrijft CHETAH, een algoritme dat automatisch de celtypen kan herkennen in genexpressie data en gezonde cellen van tumorcellen kan onderscheiden. In hoofdstuk 8 wordt onder andere CHETAH gebruikt om de genexpressie profielen van gezonde en kankercellen in HL te bestuderen. Hierdoor worden genen gevonden die actiever zijn in de kankercellen dan in de gezonde cellen en daardoor mogelijk gebruikt kunnen worden om nieuwe therapieën te ontwikkelen. Ook worden genen beschreven die essentieel zijn voor het communiceren tussen de kankercellen en de cellen in de rest van de tumor. Mogelijk kunnen nieuwe therapieën deze communicatie verstoren en daardoor de tumorcellen doden.

Dit proefschrift beschreef moleculaire studies die chemotherapiemedicijnen en nucleotideanalogen identificeerden die het DNA van gezonde cellen beschadigen, en die genen beschreven die mogelijke nieuwe doelwitten kunnen zijn voor het ontwikkelen van minder schadelijke therapieën. Deze studie en andere vergelijkbare studies verbeteren ons begrip van de schadelijkheid van kanker therapie en kunnen gebruikt worden voor het effectief ontwerpen van toekomstige klinische studies die als doel hebben om late effecten bij kinderen met kanker te verminderen. Deze kunnen bijvoorbeeld minder schadelijke therapieën ontwikkelen of testen, of de dosis van, bijvoorbeeld, platina bevattende medicijnen of ganciclovir verminderen.

List of Publications

Eline J.M. Bertrums* , Axel K.M. Rosendahl Huber*, **Jurrian K. de Kanter*,** Arianne M. Brandsma, Anaïs J.C.N. van Leeuwen, Mark Verheul, Marry M. van den Heuvel-Eibrink, Rurika Oka, Markus J. van Roosmalen, Hester A. de Groot-Kruseman, C. Michel Zwaa, Bianca F. Goemans, Ruben van Boxtel. *Elevated Mutational Age in Blood of Children Treated for Cancer Contributes to Therapy-Related Myeloid Neoplasms.* Cancer Discovery 2022; 12(8):1860–1872; https://doi.org/10.1158/2159-8290.CD-22-0120

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, Ruben van Boxtel. *Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients.* Cell Stem Cell 2021; 28 (10): 1726-1739.e6; https://doi.org/10.1016/j.stem.2021.07.012

Jurrian K. de Kanter*, Philip Lijnzaad*, Tito Candelli, Thanasis Margaritis, Frank C. P. Holstege. *CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing.* Nuclei Acid Research 2019; 47(16);e95 https://doi.org/10.1093/nar/gkz543

Aura Zelco, Vanja Börjesson, **Jurrian K. de Kanter**, Cristina Lebrero-Fernandez, Volker M. Lauschke, Eridan Rocha-Ferreira, Gisela Nilsson, Syam Nair, Pernilla Svedin, Mats Bemark, Henrik Hagberg, Carina Mallard, Frank C.P. Holstege and Xiaoyang Wang. *Single-cell atlas reveals meningeal leukocyte heterogeneity in the developing mouse brain*. Genes & Development 2021; 35:1190-1207; https://doi.org/10.1101/gad.348190.120

Freek Manders, Arianne M. Brandsma, **Jurrian K. de Kanter**, Mark Verheul, Rurika Oka, Markus J. van Roosmalen, Bastiaan van der Roest, Arne van Hoeck, Edwin Cuppen & Ruben van Boxtel. *MutationalPatterns: the one stop shop for the analysis of mutational processes.* BMC Genomics 2022; 23(1):134; https://doi.org/10.1186/s12864-022-08357-3

Axel Rosendahl Huber*, Anaïs J.C. N. van Leeuwen*, Flavia Peci, **Jurrian K. de Kanter**, Eline J.M. Bertrums, Ruben van Boxtel. *Whole-genome sequencing and mutational analysis of human cord-blood derived stem and progenitor cells.* STAR Protocols 2022; 3(2):101361; https://doi.org/10.1016/j.xpro.2022.101361

Anne Margriet Heijink*, Colin Stok*, David Porubsky, Eleni Maria Manolika, **Jurrian K. de Kanter**, Yannick P. Kok, Marieke Everts, H. Rudolf de Boer, Anastasia Audrey, Femke J. Bakker, Elles Wierenga, Marcel Tijsterman, Victor Guryev, Diana C. J. Spierings, Puck Knipscheer, Ruben van Boxtel, Arnab Ray Chaudhuri, Peter M. Lansdorp & Marcel A. T. M. van Vugt. *Sister chromatid exchanges induced by perturbed replication can form independently of BRCA1, BRCA2 and RAD51.* Nature Communications 2022; 13(1):6722; https://doi.org/10.1038/s41467-022-34519-8

Curriculum Vitae

Jurrian Kornelis de Kanter was born on October 5th 1995 in Leiden, the Netherlands. In 2013 he obtained his high school degree cum laude at Stedelijk Gymnasium Leiden. In the same year he started his bachelor's degree in biomedical sciences at Utrecht University, where his interests in genetics and cancer were sparked. He graduated in 2016 and one year later he continued his education in the master program Cancer, Stem Cells and Developmental biology (CS&D) at Utrecht University. He became acquainted with the field of bioinformatics during his first internship in the lab of Frank Holstege, under supervision of Philip Lijnzaad. Here, he developed a single-cell RNA-sequencing cell type classifier algorithm in collaborations with his supervisors. This work resulted in his first published paper. He further developed his bioinformatics and research abilities in the lab of Marcel Kool, under supervision of Sander Lambo. Here, he studied the epigenetic landscape of pediatric embryonal tumor with multi-layered rosettes (ETMR). Wanting to continue in bioinformatics of pediatric cancer, he started his PhD trajectory in the lab of Frank Holstege, and soon transitioned to the group of Ruben van Boxtel. Here, he studied the genomic consequences of pediatric cancer treatment and the microenvironment of pediatric Hodgkin Lymphoma. The results of these studies are presented in this thesis. For his first two publications, he won the "CS&D PhD Publication of the Year" award in 2022. In March 2024, he continued his career as Data Scientist at Genmab.

Acknowledgements

This thesis would not have been possible without the work, help, and support of many people. Here, I would like to thank all of you!

Allereerst, en uiteraard het meest belangrijk, waren de patiënten en ouders die hebben meegewerkt aan het onderzoek in het Prinses Máxima Centrum door hun gegevens en materiaal te delen. Dit onderzoek gaat natuurlijk om het verbeteren van de kwaliteit van leven van de patiënten. Ik heb het altijd inspirerend gevonden dat het onderzoek zowel voor hen is, als door hen mogelijk gemaakt is.

Natuurlijk moet ik jou bedanken **Ruben**. Dat zal ik uiteraard doen met een portie van je eigen enthousiasme. "Wauw! Ontzettend gaaf. Je bent echt een hele goede supervisor. Zo'n supervisor kan je gelijk publiceren. We gaan meteen 10 van deze supervisors sequensen! Nee 20!". Nee, maar nu serieus. Ik kwam heel onverwachtse in jouw groep, maar ik ben erg blij dat het zo is gelopen. Je kiest de juiste mensen in je groep die samen zorgen voor een hele fijne sfeer. Je zorgde dat ik snel in deze groep werd opgenomen en meteen bij een project kon aanhaken. Ik heb veel gehad aan jouw energie, inzicht, en snelle denkvermogen. Je denkt altijd in mogelijkheden, (bijna) nooit in problemen. Dat is zeker iets wat ik de rest van mijn carrière zal meenemen!

Frank, ik moest helaas na de eerste maand van mijn PhD je groep al verlaten omdat je Managing Director werd, maar ik ben blij dat je altijd beschikbaar bent geweest voor adviezen van de zijlijn. Dit deed je ondanks dat het voor jou zware jaren geweest zijn. Ik neem een diepe buiging voor je doorzettingsvermogen. Zodra het iets beter met je ging, kregen ik en andere PhDs meteen een mailtje: "laat het weten als je iets nodig hebt, ik ben altijd beschikbaar". Want, zoals je zelf zei: "Natuurlijk wil ik werken. Als ik nu opeens niet meer zou willen werken, maar alleen tuinieren, zou dat betekenen dat ik jarenlang iets gedaan heb wat ik niet leuk vind.". Ik hoop dat ik over 30 jaar hetzelfde kan zeggen.

Bedankt **Wouter** en **Jayne** voor jullie adviezen tijdens de jaarlijkse PhD voortgangsbijeenkomsten en voor het carrièreadvies. Ik wil jullie en **Edwin Cuppen, Lude Franke**, en **Josef Vormoor** bedanken voor het lezen en beoordelen van mijn thesis, en **José Borghans, Marcel Kool,** en **Berend Snel** voor het deelnemen aan de promotiecommissie .

Alex, or should I say fit.steem? Thank you for being my paranymph. It was a pleasure to introduce you to the lab and to work with you on the lymphoma projects. You are an extremely social person, and we got along straight away when you joined the group. I have always appreciated your Cypriot hospitality. Brunch, pastitsio, beers, mimosas, we've had everything at your place! Thanks for all the fun during paintball,

movies, the retreats, borrels, trips, and countless other occasions.

Diego "the beast" Montiel Gonzalez, thank you for being my desk buddy for the last few years. I have enjoyed our little corner of the office… although maybe you didn't think so in the beginning when I was looking at my screen and meanwhile listening to you (most of the time). Thanks for all the jokes, your endless positive energy, and all the struggles we have shared, boring programming tasks, packages not installing, and too many more to mention.

Niels, ik weet dat je niet van complimenten houdt, maar toch ga ik het nog eens zeggen: je bent echt ontzettend goed in wat je doet! Misschien wel een van de leukste dingen in afgelopen jaren vond ik om de groep binnen te komen met het Hodgkin project, daarna jouw hulp te krijgen, om vervolgens alleen nog van de zijlijn suggesties te hoeven geven terwijl jij het hele project op je nam. Het Hodgkin project heeft misschien drie jaar voor frustraties gezorgd bij jou en mij, maar ik ben echt blij dat je het op de valreep toch voor elkaar hebt gekregen om de techniek werkend te krijgen! Zet daarnaast alle borrels, workouts, en uitjes en er valt wat mij betreft geen betere projectbuddy te wensen.

It goes without saying that I've enjoyed my time with the van Boxtel buddies tremendously. Thanks all current and former members for making my PhD so much fun. I believe that there are only very few research groups that have such a good atmosphere and in which so much fun stuff is organized. For example, during covid (semi-)lockdowns, the outside-borrels at Café Koekkoeksplein were more than welcome, thank you **Eline** for hosting. Ik ben blij dat we zo'n goede samenwerking hebben gehad en elkaar goed konden aanvullen. Ik ken weinig mensen met zo veel energie, en zo'n werkmentaliteit als jij! **Axel**, ik ben dankbaar dat ik in jouw voetsporten heb mogen treden. Bijna elk project waar ik aan meegewerkt heb, gebruikte wel een techniek die jij met anderen in de groep hebt opgezet. Ik vind het leuk dat we elkaar nog met regelmaat tegenkomen, of het nu op een congres is of als tijdelijke buren! **Flavia**, thank you for being true to yourself and for always saying what you think. This led to some good and fun discussions. Best of luck, wherever you may settle down next! **Freek**, bedankt voor je kritische blik, je scherpe analyses, en je oog voor detail. In veel gevallen hebben jouw opmerkingen mijn werk echt verbeterd. Maar vooral bedankt voor het delen van de passie voor klimmen. Samen vinden we bijna altijd de beta (of een twijfelachtige andere aanpak). **Lucca**, bedankt voor de introductie tot het bolderen. Je bent ontzettend sterk, op meerdere vlakken, daar heb ik echt diep respect voor. **Laurianne**, thank you for all the hard work in the lab and for the interesting analytical discussions over a beer (or a few more). **Mark V.**, ik heb genoten van je nuchterheid en oprechtheid. Helaas treed ik langzaam toe bij de "club-van-de-twee-daagse-katers". We spreken elkaar vast snel weer daar. Over oprechtheid gesproken, bedankt voor al je oprechte nieuwsgierigheid **Anaïs**, je wist altijd met onverwachte vragen te komen. Ik zal een voorbeeld aan je nemen en de

komende jaren mijn best doen om al mijn top-5 lijstjes scherper op een rij te zetten! **Joske**, bedankt voor alle goede gesprekken die we gevoerd hebben, soms achter ons bureau, soms bij een kampvuur in België. Nu loopt jouw pad helemaal naar, en door, Amerika. Ergens ben ik jaloers, maar ik doe het je niet na. Veel plezier en succes bij de PCT, gelukkig heb je de allerlichtse en mooiste spullen die er zijn! **Mirjam**, het was me een eer om de eerste paper tijdens mijn PhD met je te publiceren. Ik ben erg blij dat ik mocht inspringen tijdens het project, en met zo'n uitkomst. Het is bewonderenswaardig hoe je de afgelopen jaren praktiserend gespecialiseerd arts zijnde, ook je hele eigen onderzoeksgroep hebt opgezet, ook al was dat niet altijd makkelijk. Volgens mij gaan daar nog heel veel mooie dingen uit voortkomen! **Friederike**, je bent vanaf het begin af aan al betrokken geweest bij mijn onderzoek. Nu je veel tijd hebt voor onderzoek is extra duidelijk te zien dat dit echt je passie is. Je bent een echte connector en initiator. Ongelofelijk bij hoeveel projecten je betrokken bent. Jij maakt de brug tussen kliniek en fundamenteel onderzoek eigenhandig waarheid, precies waar het Máxima voor staat. **Arianne**, fijn dat jij er was om me op weg te helpen in het begin van mijn PhD, ondanks het feit dat ik niet altijd even punctueel was. Jij hebt echt een vorm gegeven aan het Hodgkin project, dit had ik nooit zonder jou kunnen doen. Ik hoop dat je snel in de kliniek aan de slag mag om echt te kunnen doen wat je leuk vindt. **Markus**, bedankt voor al je bioinformatica hulp, je bent echt van alle markten thuis, niets wat je niet op kan aanpakken! **Rico**, met jou valt er altijd iets te kletsen, bedankt voor alle gesprekken, of het nu over huizen, eten, of politiek ging, met jou is het nooit saai. P.S. Sterkte met mijn code. **Vera**, wat leuk dat je bij ons in de groep bent gekomen. Je lag vanaf moment één goed in de groep. Hoe kan het ook anders met zulke leuke Driedaagse Feesten. Bedankt **Maarten**, voor de inzichten tijdens de lunchgesprekken, bijvoorbeeld in de andere soorten werk die mogelijk zijn binnen de wetenschap, en natuurlijk in garnalen. Jouw skills en kennis helpen de groep echt en het is cool om te zien hoeveel projecten van de grond zijn gekomen sinds je bent gekomen. Bedankt **Siem** en **Kees** dat ik jullie heb mogen begeleiden in jullie stages. Ik hoop dat jullie er veel van geleerd hebben, ik in ieder geval zeker. Bedankt **Annemarie** voor alle hulp en sugessties tijdens de 1-to1's. **Sophie**, **Liza,** and **Diana**, good luck with your PhDs! Finally, a big thank you to all other current and former members of the van Boxtel group, **Annina, Andrea, Damon, Emma, Georgiana, Inge, Jip, Karlijn, Marta, Madalena, Rurika, Sjors, Sofia, Sophia, Steven, Suzanne, Teuntje**, and all others!

I also want to thank all member of the Holstege group, **Aleksandra, Eduard, Ewa, Jeff, Lindy, Marian, Marit, Mariël, Michael, Philip, Thanasis, Tito**, **Tomasz,** and **Wim** for your advice, talks, and fun during my first internship and the first month of my PhD. A special thanks to you **Thanasis**, for handing over the Hodgkin project that you put so much time in. I highly appreciated all your advice over the years, pointing me in the right direction whether during my Master, PhD, or when I was looking for a job. I have a lot of respect for the way you manage the facility and, in the meantime, are a mentor to many people. Thank you, **Aleksandra** and **Eduard**

for all the help with the single-cell sequencing. **Philip**, thank you for introducing me to the wonderful world of bioinformatics and being such a patient teacher. It was an honor to work together and be co-authors on the manuscript that would become my first published article. It's due to you that I always go ToTheMax. **Tito**, where you are, there is an interesting conversation. No topic is too wild or too boring, or at least it does not influence the passion with which you talk about it. Thank you for your honest opinion and endless knowledge, whether on English grammar or D&D rules. On that note, **Francisco, Michael, Thanasis**, and **Tito**, thanks for all the evenings of Divine Smite, demons and devils, and Wild Shape. **Wim** and **Marit**, thank you for all board games and fun times during and after my internship. I am happy that we still meet up over board games and beer!

Thanks **Anna** and **Francesco** for enduring the endless conversations of our group and for the chats in the office and at the many retreats and Masterclasses. Thanks for teaching me some tricks on the slopes **Francesco**. I had a lovely time in France (and at the Maxima) with all of you, **Ravian, Britt, Raphael, Rijndert, Emma, Mieke, Irene, Alex, Lucca**, and all others. Irene, thanks for the company and conversations in Heidelberg. Thanks to **Arianna, Charlotte, Camilla, Carla, Dilara, Francisco, Guilia, Irene, Jarno, Jiayou, Juliane, Kim, Lars, Marjolein, Maroussia, Nadia, Terezinha, Sofia, Yvonne**, and all the other members of the Drost group for great (secret) retreats.

Besides all these amazing people, I have been lucky enough to meet many more people at the Maxima from many different groups and places. Thank you for all the fun times **Bas, Cedric, Chris, Evelyn, Geerte, Joanna, Konradin, Luuk, Margrit, Trisha**, and many, many others.

Apfrida, **Cindy**, **Jorian**, **Lars**, **Nienke**, and **Ronja**, it was a joy to organize the CSND retreat with all of you! I had an amazing time during the retreat with all the drinks, games, and live band (and of course great scientific content).

Awooo to the members of Kállos Sthenos, **Tito, Konstantinos, Niels**, and of course the leader **Diego** for the manly man workouts, in the dark, in the rain, and in the snow. Beast on!

Bedankt **Marijn** en **Liset** voor de samenwerking, en alle informatie die jullie telkens binnen de kortste keren wisten op te rakelen tussen al jullie andere taken door. Bedankt **Hester** voor het opzoeken van grote hoeveelheiden klinishe data. Bedankt **Ravian** voor het imaginen en de mooie plaatjes als eindresultaat. I appreciated all the discussions, input, and insightful comments from all bioinformaticians from the Maxima, and all attendees of the Bioinformatic Meetings, thank you. I also had the privilege to have collaborated with **Marcel, Lauren, Xiaoyang, Aura** and others from outside of the Maxima.

Bedankt **Vinnie** en **Floor** voor de introductie tot mijn studententijd, met een biertje in de hand darten in het minuscule gangetje, hopend dat je degene die zich de douche-kast uitwurmde niet zou raken. Mooi dat we elkaar nog steeds spreken, of het nu in Den Haag, Amsterdam, of Utrecht is.

Warner en **Ruben**, bedankt voor alle (blind geproefde) speciaalbiertjes, de strandwandelingen, een mooi huwelijk, en de studie, werk, en vooral levensreflecties in Leiden, Utrecht, Sassenheim, Den Haag, en Wageningen van de afgelopen anderhalve decennium.

Lieve Kleiers, **Wieke**, **Olivier**, **Marjolein**, en **Geart**. Wat hebben wij samen veel meegemaakt in de afgelopen jaren. Er zijn geen andere vrienden met wie ik zoveel hoogte- en dieptepunten heb gedeeld als met jullie. Dat we er allemaal voor elkaar zijn staat vast! Wat hebben we lekker gehanst op laaionde festivals, bijzonder interessante oud- en nieuwfeestjes, fietsend door heen Nederland, en op weekendjes weg. Nu is het tijd voor nieuwe hoofdstukken in jullie leven, deels aan de andere kant van de wereld, en deels hier, met de kleine Hans, ehm, Mara. Ik kijk er naar uit om deze en nog heel veel andere hoofdstukken met jullie te delen.

Bedankt **Bas**, voor de afgelopen 11 jaar vriendschap in al zijn vormen, van samen roeien, tot roeiploegjes coachen, van met je vastgezette been omhoog kratten bier wegtanken in ons aftandse huisje in Zuilen tot uren op de racefiets, van de vele avonden in de Utrechtse kroegen, tot kilometers wandelen, van totaal onverantwoord slingerend op de fiets naar huis, naar Oscar vasthouden op jouw bank. Maar een ding was constant: de goede gesprekken over mensen, studie, en nu werk en kinderen, maar altijd met een stevige vleug scepsis en sarcasme.

Bedankt **Rob** en **Jolanda** voor de hartelijkheid waarmee jullie mij altijd verwelkomd hebben. De deur stond altijd open, of het nu voor een koffie was, een kerstdiner, of een weekje werken tijdens lockdowns. Bedankt ook **Martijn** en **Jasper**, voor alle potjes Mario Party/Kart, Ticket to Ride, en Monopoly.

Lieve **pap** en **mam**, er was niemand geïnteresseerder in de inhoud van mijn werk dan jullie. Altijd luisterde jullie vol interesse (soms met aantekeningenboekje erbij!) of het nu ging over de kleine details van de biologie of de algemene uitdagingen en frustraties die een PhD met zich meebrengt. Bedankt voor alle liefde en vertrouwen die jullie me hebben gegeven, vroeger en nu nog steeds. Bedankt **Floor**, voor alle discussies die we gevoerd hebben over de jaren heen. Vroeger wat feller, tegenwoordig wat liefdevoller. Samen gingen we door het PhD traject heen, en het was fijn om dat samen te kunnen analyseren, en soms bij elkaar wat (of heel veel) te kunnen luchten. **Iris**, bedankt voor alle humor die we samen kunnen delen, van oude internet referenties tot Friends of Ronald Goedemondt. Ook al begrijpt soms niemand anders het, met jou heb ik zo vaak in een deuk gelegen. Bedankt **Rik** voor je mooie verhalen en inzichten, en bedankt **Job** alle spelletjes en de onuitputbare hunebedden anekdotes.

Lieve **Chantal**, bedankt dat je me altijd hebt gesteund de afgelopen jaren. Een knuffel als het even zwaar was, koken als ik weer eens een aantal avonden achter elkaar moest doorwerken, het aanhoren van mijn geneuzel over saaie informatica problemen, en het analyseren van werksituaties, je stond altijd voor mij klaar. Fijn dat er altijd (lekker lang van tevoren dankzij jou) weer een vakantie, een sauna weekendje, of een uitje was waar we samen naar uit konden kijken. Ik ben heel blij dat we een mooi plekje in Utrecht hebben waar we samen echt een thuis van hebben gemaakt. Ik hou van je en ik kijk uit naar wat de toekomst nog te brengen heeft voor ons samen.