HHS Public Access

Author manuscript

Cancer Discov. Author manuscript; available in PMC 2023 October 03.

Published in final edited form as:

Cancer Discov. 2023 April 03; 13(4): 844–857. doi:10.1158/2159-8290.CD-22-0956.

Dynamics of age- versus therapy-related clonal hematopoiesis in long-term survivors of pediatric cancer

Kohei Hagiwara[1], **Sivaraman Natarajan**[1,*], **Zhaoming Wang**[2,*], **Haseeb Zubair**[1,*], **Heather L. Mulder**[1], **Li Dong**[1], **Emily M. Plyler**[1], **Padma Thimmaiah**[1], **Xiaotu Ma**[1], **Kristen K. Ness**[2], **Zhenghong Li**[2], **Daniel A. Mulrooney**[2,3], **Carmen L. Wilson**[2], **Yutaka Yasui**[2], **Melissa M. Hudson**[2,3], **John Easton**[1,#], **Leslie L. Robison**[2,#], **Jinghui Zhang**[1,#]

[1]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA

[2]Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, Tennessee, USA

[3]Department of Oncology, St. Jude Children's Research Hospital, Memphis, Tennessee, USA

Abstract

We present the first comprehensive investigation of clonal hematopoiesis (CH) in 2,860 long-term survivors of pediatric cancer with a median follow-up time of 23.5 years. Deep-sequencing over 39 CH-related genes reveals mutations in 15% of the survivors, significantly higher than the 8.5% in 324 community controls. CH in survivors is associated with exposures to alkylating agents, radiation, and bleomycin. Therapy-related CH shows significant enrichment in *STAT3*, characterized as a CH-gene specific to Hodgkin lymphoma survivors, and *TP53*. Single-cell profiling of peripheral blood samples revealed *STAT3* mutations predominantly present in T-cells and contributed by SBS25, a mutational signature associated with procarbazine exposure. Serial-sample tracking reveals that larger clone size is a predictor for future expansion of age-related CH clones, while therapy-related CH remains stable decades post-treatment. These data depict the distinct dynamics of these CH subtypes and support the need for longitudinal monitoring to determine the potential contribution to late effects.

Keywords

Clonal hematopoiesis; Late effects; Cancer survivorship; Pediatric cancers; *STAT3*

[#]Correspondence: John Easton, Department of Computational Biology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, MS1135, Memphis, TN 38105, Phone: (901) 595-7359, john.easton@stjude.org; Leslie L Robison, Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, 262 Danny Thomas Place, MS735, Memphis, TN 38105, Phone: (901) 595-5817, les.robison@stjude.org; Jinghui Zhang, Department of Computational Biology, St. Jude Children's Research Hospital, 262 Danny Thomas Place, MS1135, Memphis, TN 38105, Phone: (901) 595-5935, jinghui.zhang@stjude.org.
[*]These authors contributed equally.

CONFLICTS OF INTERESTS

None declared.

INTRODUCTION

Clonal hematopoiesis (CH) refers to the presence of hematopoietic stem cell (HSC) subpopulations that have become genetically distinct from the germline genome via acquisition of somatic mutations. While CH is an age-related phenomenon in the general population (1), it can also be induced by exogenous factors such as chemotherapy, which can affect both the emergence and evolution of CH clones. In addition to the higher prevalence among people previously treated for malignancy (2, 3), CH is now recognized as a precancerous state of blood cancers (4 – 7), highlighting the need for monitoring malignant transformation of secondary neoplasms for cancer survivors (1 – 3).

With the overall improvement of pediatric cancer survival rates, it has become evident that pediatric cancer survivors are at risk of accelerated physiological aging (8, 9). Despite this, the study of CH in pediatric cancer survivors has been limited by poor representation of this population in studies that focused primarily on survivors of adult-onset cancers (2, 3) or has focused on survivors with limited duration of follow-up or small cohort sizes (10 – 12). For example, a recent study by Bertrums *et al.* (12) showed that chemotherapy can induce mutations in hematopoietic stem cells directly or indirectly by studying 24 pediatric patients with a median follow-up time of 1.65 years (range: 0.2 – 10 years). These studies have greatly improved our understanding of the acute landscape of therapy-related CH. However, as life expectancy of long-term (>5 years) survivors of pediatric cancer can exceed 50 years after treatment (13), the long-term profile of CH in survivors, including the evolutionary trajectory during their lifespan, may inform the design of clinical management of late effects.

To fill this knowledge gap, we analyzed CH in the St. Jude Lifetime Cohort study (SJLIFE), a retrospective cohort with prospective clinical follow-up of pediatric cancer survivors treated at St. Jude Children's Research Hospital since 1962 (14, 15). We performed deep-sequencing on DNA extracted from peripheral blood samples obtained from 2,860 SJLIFE survivors (age range: 6.0 – 66.4 years, median 31.6) to enable comprehensive CH analysis in this relatively young population. Furthermore, we performed statistical modeling to probabilistically assign CH into age- and therapy-related subtypes, in order to characterize their distinct dynamics in clonal expansion.

RESULTS

Higher CH incidence in pediatric cancer survivors

To estimate the prevalence of CH in pediatric cancer survivors, we performed targeted sequencing of 39 hematological malignancy-associated or cancer predisposing genes (Supplementary Table S1) using peripheral blood samples from 2,860 SJLIFE survivors and 324 community controls (Fig. 1A). The survivors were followed up over 5.1 – 51.1 years (median 23.5 years) from cancer diagnosis and were aged 6.0 – 66.4 years (median 31.6 years) at the time of sample collection, which was slightly younger than the community controls (18.3 – 70.2 years, median 34.6) (Supplementary Figure S1). The childhood cancer diagnoses of the survivors included leukemias (35%), lymphomas (19%), central nervous system malignancies (CNS, 11%), and non-CNS solid tumors (35%). Given the

relatively young age of our cohort, the median raw sequencing depth was set at 15,987× (1,720× after de-duplication) so that we could detect CH clones present at a variant allele frequency (VAF) as low as 0.1%. To reduce the error rate inherent in deep sequencing data, additional analyses such as computational error suppression (16), outlier detection (17) adjusted for sequencing context (Fig. 1B), and indel realignment (18) were employed in variant analysis (Methods). Approximately 40% of the putative variants were subjected to orthogonal validation by digital droplet PCR (ddPCR), while the validity of the remaining variants was inferred by modelling the site-specific background error (Fig. 1A, Methods, and Supplementary Methods). Altogether, we identified 540 validated CH variants with a median VAF of 0.4% (range 0.1 – 29.5%) for further analysis (Supplementary Table S2).

Higher CH prevalence was found in the survivors compared to the community controls in each age category (Fig. 1C) with a statistically significant increase in the overall prevalence: 15.0% of survivors (95% confidence interval [CI]: 13.7 – 16.3%) vs. 8.6% of controls (95% CI: 5.6 – 11.7) (Fisher's exact $p = 1.44 \times 10^{-3}$). When limited to variants with VAF ≥ 2%, a threshold commonly used for other studies, CH prevalence was 1.99% (57/2,860) and 0.93% (3/324) for survivors and controls, respectively. Notably, at this cutoff, the 1.99% prevalence in our relatively young survivorship cohort (median of 31.6 years) is comparable to the prevalence in a general population aged 50 – 59 years (2.54%, 138/5,441, Fisher's exact $p = 0.128$) (5). As hematopoietic stem cells acquire mutations during cell division, we compared the leukocyte telomere length, a biomarker for cell division, in CH-positive survivors and controls using the accompanying whole-genome sequence (WGS) dataset (19, 20). The shortening of telomere with age relationship represented by least-square regression fit showed nearly parallel lines (Figure 1D). This indicates that the telomere attrition rates, which reflect the rates of underlying hematopoietic stem cell (HSC) division, were very similar between the controls (41.4 bp/year) and survivors (45.1) ($p = 0.413$). This suggests that the CH variants were acquired at a similar rate in survivors and controls. Therefore, the increased CH prevalence in the survivors was likely related to cancer and/or treatment-related exposures in childhood rather than an acceleration of cell division.

Associations of CH status with therapeutic variables

To investigate the origin of elevated CH in pediatric cancer survivors, we analyzed the association of CH with demographic variables and cancer treatment exposures. While no association was found with sex and race/ethnicity, the prevalence of CH was correlated with age in both survivors and controls as expected (Supplementary Table S3). We next sought to evaluate association of specific therapies with CH, as cancer treatments typically involve multiple therapeutic agents and modalities. In multivariable logistic regression analysis, after adjusting for age at sample acquisition and age at diagnosis, CH was associated with alkylating agents, bleomycin, and estimated RT dose to active bone marrow (Fig. 2A). Specifically, exposures to alkylating agents and RT were independently associated with CH in a dose-dependent manner, which was not observed for bleomycin (Fig. 2B). As treatment differs by cancer type, we analyzed the association of CH prevalence by diagnosis (Fig. 2C). After adjusting for age, survivors of Hodgkin lymphoma, soft tissue sarcomas, germ cell tumor, rhabdomyosarcoma, neuroblastoma, non-Hodgkin lymphoma, or acute lymphoblastic leukemia were more likely to develop CH compared to the controls; these

diagnoses were collectively represented as *Dx** (Fig. 2C). We also found that a significantly higher proportion of survivors with *Dx** had been exposed to alkylating agents (all dose tertiles, details in Methods) than the other diagnoses (labeled *Dx* in Fig. 2D) (67.0% vs. 35.9%, proportion test $p = 1.83 \times 10^{-57}$). Bleomycin was administered only to a small number of cases (n = 147) with the 2nd tertile exhibiting a significantly higher proportion in cancer diagnoses significantly associated with CH risk (2.47% vs 0.20%, $p = 7.18 \times 10^{-6}$). The 3rd tertile of bleomycin dose was primarily among survivors of osteosarcoma, a cancer type that barely missed the cutoff for *Dx** (Fig. 2C), which led to the statistically insignificant finding in this category (Supplementary Figure S2). There was no difference in the distribution of RT 3rd tertile (18.3% vs 17.2%, $p = 0.477$). However, in a subset analysis of survivors treated without chemotherapy (n = 475, Supplementary Figure S3), where a potential effect-masking by chemotherapy was uncovered, we similarly observed a significantly higher proportion of the 3rd tertile (30.9% vs 16.1%, $p = 2.09 \times 10^{-5}$).

Although exposures to alkylating agents, bleomycin, and RT were associated with elevated CH prevalence in the survivors, there may exist additional contributing factors. To address this hypothesis, we performed a mediation analysis to quantify how much of the CH association with cancer type could be explained by these three therapies (21). Exposure to the statistically significant tertiles of alkylating agents, bleomycin, or RT was defined as a dichotomous mediator in the analysis. After adjusting for ages at diagnosis and at sample collection, the therapy mediator explained 74% of the association between CH and cancer diagnosis, confirming that the CH development in our cohort was largely contributed by these therapies.

CH association with germline pathogenic mutations in survivors

In addition to the therapy variables, germline mutations may have also played a modifier role in CH development. We first analyzed the CH association with 112 germline pathogenic mutations identified in SJLIFE cohort, which were previously classified based on the American College of Medical Genetics and Genomics (ACMG) guidelines in 60 cancer predisposition genes (CPG) known to be associated with autosomal dominant cancer predisposition syndromes with moderate to high penetrance (20) (Supplementary Table S4). The CH prevalence was 8.9% in carriers of CPG mutations vs. 15.3% in non-carriers, respectively (Supplementary Table S4). The reduced prevalence in the carriers, albeit statistically insignificant (Fisher's exact $p = 0.078$), likely reflected the fact that more than 50% of such mutations occurred in survivors of cancer types with the lowest CH incidence due to less intensive therapy (Fig. 2C), i.e., retinoblastoma (n = 35), Wilms tumor (n = 9), or CNS malignancies (n = 24).

We also studied the effect of deficiency in DNA Damage Repair (DDR) on CH prevalence (Supplementary Table S4) using germline pathogenic mutations in 127 DDR genes selected by our prior study (22). The overall CH prevalence, 17.9% in carriers vs. 14.9% in non-carriers (Fisher's exact $p = 0.417$), was not affected by germline mutation status. However, when stratifying the survivors by therapy exposure, non-irradiated survivors (n = 1,359) had a significantly higher CH prevalence in mutation carriers (21.2%) than in non-carriers

(11.7%, Fisher's exact $p = 0.0496$) indicating that RT might have masked the pathogenic effect of mutations in irradiated cases ($p = 0.730$).

Molecular characteristics of age- versus therapy-related CH subtypes

We proposed an additive model for CH clones detected in survivors, based on the association of therapy with CH prevalence (Fig. 2) and a cell division rate comparable between survivors and community controls in this cohort (Fig. 1D). Under this model, CH in survivors is an admixture of therapy- and age-related clones. Age-related CH clones in the survivors are generated at a rate similar to those in the community controls over time (Supplementary Figure S4). As the vast majority of CH-positive survivors (85% or 365/430) harbored only a single CH event, we classified the clones into age- or therapy-related categories by the following two steps. First, the probability of developing age-related CH in a given age category was estimated by logistic regression using the control data. Second, the probability of developing CH in a given age category and by therapy exposure was estimated in the survivor cohort. In the second regression, the age effect estimated by the first regression was included as an offset term so that the age impact on a survivor and a control will be the same at a given age category as in the proposed model (Supplementary Figure S4) (Methods). Using these two models, we computed probabilities of developing CH for each survivor, given their age and treatment exposures, and when the estimated contribution from therapy was higher than that of age, the CH of the sample was designated as *inferred therapy-related* (iTR), otherwise as *inferred age-related* (iAR). Of the 430 CH positive survivors, this approach found 214 samples as iAR, which accounted for 7.5% in the survivor cohort, comparable to 8.6% CH prevalence in the control (Fisher's exact $p = 0.439$). Prevalence of iAR in the survivors increased with age, matching the course of the community controls (Fig. 3A).

To provide independent validation for this model, we first compared the mutation spectra of iAR and iTR, which reflects the underlying mutagenesis process and subsequent selection. As shown in Fig. 3B, the spectra from the control and iAR were nearly identical, exhibiting dominance of the cytosine (C)>thymine(T) transition in 63.0% and 65.8% of the substitutions respectively, consistent with expected age-related processes (5). By contrast, the iTR spectrum was different from iAR in that C>T transition accounted for only 45.7% of substitutions (post hoc $p = 2.05 \times 10^{-3}$). While samples with four CH mutations were only observed in iTR, the distribution of CH mutation burden did not differ across the control, iAR, and iTR (Kruskal–Wallis $p = 0.454$, Fig. 3C). Next, we examined the mutation frequency in *DNMT3A, KRAS, TP53, TET3*, and *STAT3*, the five most frequently mutated genes in both survivors and controls (Fig. 3D) These five genes had similar mutation frequencies in control and iAR samples of the survivors, consistent with the iAR classification. By contrast, the frequencies in iTR showed considerable differences —significantly higher frequency was found for *STAT3* (17.9% in iTR vs. 6.25% in controls, FDR $q = 2.61 \times 10^{-6}$) and *TP53* (17.5% vs. 12.5%, $q = 0.033$) and a lower rate was found for *DNMT3A* (19.0% vs 28.1%, $q = 2.48 \times 10^{-5}$). While elevated *TP53* mutation frequency in therapy-related CH is consistent with previous studies that reported *TP53* as a therapy-related CH gene (2, 3, 23), *STAT3* has not been characterized as therapy-related.

Therapy-related STAT3 mutations in survivors of Hodgkin Lymphoma (HL)

Identification of *STAT3* as a frequently mutated CH gene is a new finding in our pediatric cancer survivorship cohort. Notably, all *STAT3* mutations were in the Src homology2 domain (Fig. 3E), which contains mutation hotpots in hematological and lymphoid cancers. The most frequent hotspot mutation, *Y640F*, was also the predominant mutation in our cohort present in 36 cases. Interestingly, *STAT3* mutations were significantly enriched in survivors of Hodgkin lymphoma (9.3% in HL survivors vs. 1.4% in survivors of pediatric cancers other than HL, Fisher's exact $p = 9.21 \times 10^{-14}$, Supplementary Table S5). This suggests that *STAT3* is a HL-specific CH gene, which was further supported by the lack of enrichment of *STAT3* mutations in survivors of cancers other than HL (1.4% in other cancers vs. 0.6% in controls, $p = 0.426$).

For three survivors with *STAT3 Y640F*, we sought to identify the mutant cell phenotype by the Mission Bio Tapestri assay, which jointly profiles the genotype and phenotype in the same single cell using oligo-conjugated antibodies (Methods). While *STAT3 Y640F* was detected in all blood cell types, it was highly enriched in T cells across all three cases (Fig. 3F, Supplementary Figure S5). Among the subtypes of T cells, the CD8+ populations consistently harbored this mutation (Supplementary Figure S5). This suggests that the mutant T-cells may have accelerated proliferation, as *STAT3 Y640F* has previously been identified as a gain-of-function driver mutation in T-cell large granular lymphocyte leukemia, a lymphoproliferative disorder characterized by CD3+ CD8+ T cell expansion (24). The broad spectrum of mutation-positive cell types coupled with its predominance in T cells argues against the possibility that *Y640F* arise from residual HL, as classical HL is of B cell origin (25, 26).

To evaluate whether mutagenic processes related to therapy might explain the association between *STAT3* mutation and HL, we performed single cell WGS analysis on mutant and wildtype cells in HL survivors (Methods) and were able to analyze 4 cells for each genotype from SJHL018702, an iTR case whose blood sample collected 22 years after the initial HL diagnosis was used for this purpose (Fig. 3G). Somatic SNVs were identified by using the bulk WGS data as the matching control (Methods). Compared to the expected mutation burden of 1,000 – 2,000 for a normal peripheral blood cell aged 37.7 years (27), all cells had elevated mutation burden. The mutation burden of *STAT3 Y640F* mutant cells was 3-4 times higher than those of the wild-type cells (Fig. 3G, middle). Mutation signature analysis of these cells only extracted COSMIC SBS25, a post-chemotherapy signature specific to HL cell lines (28) or normal tissues of HL survivors (29, 30), at cosine similarity of 0.91. Multiple signatures, including COSMIC SBS1, 5, 11, and 25 with relative contribution of 2.4%, 24.5%, 16.1% and 57.5%, respectively, at cosine similarity of 0.94, were extracted for the wild-type cells (Fig. 3G, right). SBS25 preferentially converts T to adenine (A) when C or guanine (G) precedes. Therefore, the probability that *Y640F* (GTA>GAA) was induced by SBS25 was estimated to be 0.985 to 1.0 in the mutant cells.

The near 100% probability of *Y640F* being induced by SBS25 based on mutation signature analysis of single-cell whole-genome sequencing data prompted us to examine the sequence context of all *STAT3* mutations in our cohort. Indeed, *N647I* (GTT>GAT), the only other *STAT3* mutation (besides *Y640F*) enriched in HL (Supplementary Table S5), also matches

the predominant pattern of SBS25. Recently, Santarsieri *et al.* (30) found that procarbazine, an alkylating drug used in HL treatment, is likely responsible for the SBS25 signature. In our cohort, 192 of 356 HL survivors, including SJHL018702 (Fig. 3G), were exposed to procarbazine. We found that prior exposure to procarbazine was associated with CH-positive HL survivors harboring mutations of G/C[T>A]N context, i.e., the pattern that matches SBS25 (Supplementary Table S6, Fisher's exact $p = 5.55 \times 10^{-6}$) but not those with other mutation context ($p = 0.893$), suggesting the specific contribution of procarbazine to SBS25.

Dynamic characteristics of age- versus therapy-related CH clones

To examine clonal expansion pattern, we plotted VAF, a surrogate measure for clone size, for iAR over age and for iTR over follow-up time, respectively (Fig. 4A). We tested if the pattern differed across clone sizes by performing quantile regression for the lower 25th (Q_{25th}), median (Q_{50th}), and upper 25th (Q_{75th}) quartiles of VAF values, which respectively correspond to smaller, medium, and larger clones. While iTR clones, regardless of their size quantiles, remained stable over follow up, there was a modest but significant increase of the Q^{25th} quartile in iAR over age. The iAR pattern was replicated when we analyzed two independent age-related CH datasets from previous studies using the same approach (2, 31) (Supplementary Figure S6). Encouraged by this reproducible finding, we further analyzed the correlation between clone size and growth direction using serial samples (Fig. 4B). VAF change during the two timepoints (median time interval, 4.13 years, range 0.23 – 7.9) was plotted to show the growth direction (i.e., expanding and non-expanding) of an iAR (n = 46) or iTR (n = 55) clone (Supplementary Figure S7, Supplementary Table S7); the majority of CH clones did not expand in this analysis (76% of iAR and 64% of iTR, Fisher's exact $p = 0.179$) (inset in Fig. 4B). Amongst the iAR clones, there was a tendency for the non-expanding clones to be smaller than the expanding clones at the initial time point (non-expanding and expanding clone median VAF, 0.22% vs. 1.25%, Mann–Whitney $p = 6.28 \times 10^{-4}$) (Fig. 4B), suggesting the trend at Q_{25th} was due to the loss of smaller clones. We speculate that this modest Q^{25th} trend in iAR could be a prelude to the precipitous decline in diversity of hematopoietic stem cells (HSCs) in the elderly (>70 years old) reported recently by Mitchell *et al.* (32). By contrast, such a tendency was not detected in iTR clones (0.33 vs. 0.48%, $p = 0.183$), reminiscent of the stable VAF over follow-up times across the survivorship cohort (Fig. 4A). We then tested whether these patterns could arise by performing a simulation using the branching HSC division model, which was developed to explain age-related CH dynamics (33). Indeed, simulated CH clone trajectories recapitulated the association between VAF and growth pattern of iAR (Supplementary Figure S8). To evaluate whether the contrasting pattern in iTR could be attributed to the early onset of therapy-related clones, we performed a second simulation using a modified model that generated CH trajectories driven by the same dynamics but initiated in the range of childhood cancer treatment (i.e., 0 – 20 years old) (Methods). However, this modified simulation produced a pattern similar to iAR, refuting the hypothesis that the timing of mutational acquisition led to the lack of the association between clone expansion and VAF in iTR.

The median follow-up time of the serial samples profiled in this study was 33.1 years (range 14.0 – 48.4 years), and this long follow-up time allowed us to evaluate the long-term

effect of cytotoxic therapies on the growth rate of CH clones in iTR and iAR clones. There was no significant difference in the growth rate of iAR and iTR clones harboring the most frequently mutated genes including *TP53* (Supplementary Figure S9, top panel). This indicates that there is minimal long-term effect of cytotoxic therapies on accelerating the growth of clones harboring mutations including *TP53*, a DNA damage response gene — a different effect was shown previously in samples obtained with a short follow-up time (median was 0.79 years) (3). Interestingly, iTR clones with *TP53* mutations were significantly larger than their iAR counterparts (Supplementary Figure S9, bottom panel), which can be attributed to an early clonal expansion under the selective pressure of therapy, consistent with observation by Bolton *et al.* (3) using samples with short follow-up time. Altogether, these findings suggest that the selective pressure of cancer therapy on clonal expansion may not be permanent and can be attenuated over decades.

DISCUSSION

In this first comprehensive analysis on CH of long-term survivors of pediatric cancer, we detected a higher CH prevalence in the survivor cohort compared with the community controls (Fig. 1C). The results are consistent with findings from recent studies that reported the effect of cancer therapy on CH in cohorts comprised mostly of adult cancer patients (2, 3). More importantly, the long-term follow (median 24 years) of the SJLIFE cohort enabled us to confirm that the elevated CH is a chronic condition that persists for decades, which has potential clinical relevance to inform clinical trials of early interventions in childhood cancer survivors. Our study only analyzed survivors who were alive at the cohort enrollment and thus able to provide a blood sample for sequencing. A review of all >5-year survivors at St. Jude found that only 0.68% succumbed to Secondary leukemias/MDS before they could be recruited to SJLIFE. Given the very low frequency of such events, our CH analysis provides a general representation of long-term survivors of pediatric cancer.

While radiotherapy was found related to CH development by Bolton *et al.* (3) as well as this study, the association with chemotherapeutic agents can be affected by different cancer diagnoses examined in different cohorts. For example, platinum-based agents were associated with CH in the mostly adult cohort (3) but not in our pediatric cancer survivors. Notably, leukemias and lymphomas, which account for 54% (1,557/2,860) of the cases in our cohort, were rarely (1.1% or 17/1557) treated with platinum. After excluding these hematological malignancies, multivariable logistic analysis (Fig. 2A) showed a marginal, albeit barely insignificant, association between CH and platinum (odds ratio: 1.14, 95% CI: 0.996 – 1.31 with p = 0.057). By contrast, platinum-based agents were the most frequently used drug in Bolton *et al.* (3) with 44% of the patients being exposed (5,236 exposed vs. 6,695 unexposed), rendering sufficient statistical power to detect such an association.

HSC turnover rates of survivors versus controls were comparable as indicated by their near identical telomere attrition rate (Fig. 1D). This provides a rationale for separating age- versus therapy-related events in the survivor cohort. A striking molecular finding in therapy-related CH is the identification of *STAT3* as a CH gene in Hodgkin Lymphoma (HL) survivors (Table S5). Single-cell profiling showed that *STAT3* mutations are predominantly present in T cells and are therefore unlikely to arise from residual disease as classical HL is

of B cell origin (25, 26). The sequence context of the two enriched *STAT3* mutations in HL, *Y640F* and *N647I*, also matches the predominant pattern of COSMIC mutation signature SBS25. The causality of SBS25 to *STAT3 Y640F* was further supported, with an estimated probably of 0.985 to 1.0, by single-cell whole-genome sequencing (scWGS) data generated from a HL survivor exposed to procarbazine. Recently, Santarsieri *et al.* (30) found that procarbazine, an alkylating drug used in HL treatment, is likely responsible for SBS25 and recommended that procarbazine be replaced with dacarbazine because dacarbazine is as effective without incurring an excess mutation burden (30). Our findings from scWGS profiling as well as correlative analysis of procarbazine and CH mutation context in the HL survivors (Supplementary Table S6) provide further support on their recommendation.

While clone size estimated by VAF can predict the subsequent growth of age-related CH clones (Fig. 4B and Supplementary Figure S7), therapy-related clones appear to lack this feature, at least within the VAF range analyzed in this study (1st and 3rd quartiles, 0.2% and 0.7%). This difference in clone dynamics may be attributable to non-mutational factors, in addition to higher mutation burden suggested by the outliers in (Fig. 3C) and single cell analysis (Fig. 3G). Because of their young age at cancer diagnosis, the survivors analyzed in this study received cancer therapy when their hematopoietic and immune systems were still undergoing development (median 7.1 years old at diagnosis, range: 0.0 – 23.6), which might have modulated normal development (34) especially for survivors with CH classified as iTR. While mutated cells with relatively low fitness are routinely eliminated by immune system surveillance in healthy individuals (35, 36), the altered microenvironments in the survivors might have imposed a weaker selection on CH clones, thereby tolerating clones with various adaptation potentials including those behaving erratically.

Our study has several limitations. First, CH characterization was based on deep sequencing of a panel of 39 genes (Supplementary Table S1) and missing newly identified CH genes is an inherent problem with this approach. For example, our gene panel does not include *PPM1D* and *CHEK2*, which were recognized as prominent driver genes in therapy-related CH in recent studies (3, 23). To assess the impact of missing data, we performed targeted sequencing on *PPM1D* as the mutation frequency is twice as high as *CHEK2's* (3). By including only samples available for *PPM1D* analysis (2,185 survivors and 311 controls) (Supplementary Figure S10A and B) there is a slight increase of CH prevalence in survivors by 0.3% (from the original 15.0% to 15.3%), which does not have major impact on the main findings presented in this study. Second, CH variants in non-coding regions or in non-driver genes, which are not assayed by gene panels, can be important markers for defining the CH landscape. A recent study by Mitchell *et al.* (32) performed whole-genome sequencing on single cell-derived colonies of HSC and found that the majority of expanding HSC subpopulations are not characterized by known drivers. We do recognize, however, that with the current technology, unbiased approaches such as high-coverage (e.g., 100×) WGS or single-cell WGS as used by Mitchell *et al.* (32) is not scalable to profiling a larger cohort such as SJLIFE. Future studies may consider leveraging the strengths of both approaches: the insights gained from a global view of the entire cohort, enabled by panel sequencing, can be complemented by evolutional trajectory mapped by unbiased genome wide sequencing in selected cases.

In this study, we profiled the blood samples collected at patients' recent hospital follow-up visits to maximize the sensitivity in CH detection. Despite this, the relatively young ages of the survivors in our cohort (median 31.6 years, range: 6.0 – 66.4) may not have unveiled the full scope of adverse health conditions that will require continued follow-up efforts. Furthermore, our study has unveiled a growth profile of CH clones different from samples collected from our long-term follow-up cohort from that of samples collected closer to the time of cancer diagnosis and treatment, as in the previous study by Bolton *et al.* (3). These data support the need for longitudinal surveillance of CH in pediatric cancer survivors to evaluate its chronicity and association with clone size, a key feature used for risk stratification in other populations (1).

METHODS

Study population

Participants were enrolled in the SJLIFE study, a retrospective cohort with prospective clinical follow-up of childhood cancer survivors treated at St. Jude Children's Research Hospital (SJCRH) (37). This study was conducted in accordance with the Declaration of Helsinki and approved by the St. Jude Children's Research Hospital institutional review board. Consent for participants under 18 years of age was provided by a parent or legal guardian. All participants aged 14 years and older provided written informed consent; individuals aged between eight and 13 years provided verbal assent. Eligibility criteria and sample quality control for sequencing analysis were described previously (20) (Supplementary Methods 1). The current study included pediatric cancer patients who survived at least five years since diagnosis. A community control group (SJLIFE controls) consisting of 324 individuals without a history of pediatric cancer with frequency-matched demographic information (i.e., age, sex, and race/ethnicity) were included for comparison purposes. For individuals with blood samples collected at multiple time points, the most recent samples were used for the initial discovery phase of targeted sequencing. Demographic information and key clinical variables used for this study is presented in Supplementary Table S8.

CH variant detection

Peripheral blood samples were processed for DNA library construction via hybrid capture-based enrichment over the coding regions of 39 genes implicated in hematological malignancy or cancer predisposition (Supplementary Table S1). Libraries were sequenced on a NovaSeq6000 to a target depth of 10,000, and reads were mapped to the GRCh37 genome by BWA (Supplementary Methods 2) followed by running CleanDeepSeq (16) to remove reads with high error rate. Prior benchmark analysis based on dilution experiment showed that error suppression by CleanDeepSeq enables detection of variants at VAF of 0.05%-0.1% (16) and showed high concordance with UMI-based VAF in our recent testing on a publicly available UMI data set (38). CH variants with VAF > 0.1% were identified in the following four steps. i) Putative somatic CH SNVs were detected as outliers of alternative allele count and alternative allele fraction distribution of trinucleotide context substitutions (e.g., GCG>GTG) within each sample (Supplementary Methods 3.1). Putative somatic indels were detected by Bambino (39), which was modified for indel detection (40), with realignment

(18) (details are described in Supplementary Methods 3.2). This approach has higher sensitivity in detecting rare variants than conventional approaches used for somatic variant detection (Supplementary Methods 3.3). ii) Non-protein coding or synonymous SNVs as well as non-protein-coding indels were filtered; iii) Orthogonal validation by digital droplet PCR (ddPCR) was performed for variants in ~40% of the genes. Given the novelty of *STAT3* CH variants, we analyzed all *STAT3* variants by ddPCR provided there were sufficient DNA samples for the assay. Details are described in Supplementary Methods 4 and Supplementary Table S9. iv) For variants not selected for ddPCR or having a failed assay, their validity was inferred by building a binomial model based on read counts of the validated and not-validated read-out of ddPCR assay (Supplementary Methods 5). Also, amplicon-based sequencing on *PPM1D* was performed targeting exons 5 and 6 where somatic CH mutations are primarily found (2) and detailed in Supplementary Methods 6. The number of variant calls processed at each filtering step is shown with the CH variant detection flowchart (Supplementary Methods 7).

Germline genetic data source

Carrier status of germline mutations (20, 22) and telomere length estimation (19) were extracted from previously published data. For analyses related to cancer predisposition gene (CPG) mutations, we used only *bona-fide* pathogenic variants on 60 genes that have been associated with autosomal dominant cancer-predisposition syndromes (SJCPG60) (20, 41), thus, variants classified as “likely pathogenic” were not included in this analysis. Variants annotated as mosaic in Wang *et al.* (20) were excluded as they could be potentially CH variants. Analyses related to mutations in DNA damage repair (DDR) genes were based on Qin *et al.* (22), which analyzed 127 DDR genes from 6 pathways: homologous recombination, non-homologous end joining, nucleotide excision repair, mismatch repair, base excision repair and Fanconi anemia. Similar to CPG, we only included mutations classified as “Pathogenic” by ClinVar (version 2022-10-1).

Treatment exposure analysis

Treatment details were extracted from medical records. Chemotherapy variables were expressed as cumulative dose received per body surface area. To estimate the radiation doses relevant to CH, we estimated the maximum region-specific doses abstracted from radiation oncology records to the age-specific body distribution of active bone marrow (42). For survivors who had received autologous transplants (n = 70), we modified the therapy doses by excluding therapies undertaken between stem cell harvest and transplantation. For those who received allogenic (n = 1) or syngeneic (n = 1) transplantation, we set the therapy exposures to zero assuming complete replacement by the donor cells. The treatment data in Supplementary Table S8 were based on the modified values. To facilitate comparison of various magnitudes, all cumulative doses were scaled by standard deviation. When therapy dose was divided into tertiles, the exposed population was divided into three parts at the 1/3 and 2/3 quantiles of the ordered dose distribution so that the first, second, and third tertiles represent low, medium, and high levels of exposure, respectively. In mediation analysis, pediatric cancer diagnoses were binned into CH-enriched and not-enriched groups (Fig. 2C) represented as *Dx** (including HL, soft tissue sarcoma, germ cell tumor, rhabdomyosarcoma, neuroblastoma, non-Hodgkin lymphoma, acute lymphoblastic leukemia) and *Dx* (including

acute myeloid leukemia, osteosarcoma, Ewing sarcoma family of tumors, retinoblastoma, central nervous system malignancies, Wilms tumor, and others), respectively. This binary diagnosis was examined by defining the CH status as outcome and the exposure to alkylating agents, bleomycin, or RT with significant dose (Fig. 2B) as binary mediator. Logistic regression adjusted for age at sample collection and age at diagnosis was used to assess the mediation.

Modelling age- and therapy-related CH

Using the control data, the probability, p, to develop age-related CH was estimated at a given age-category binned as in Figure 1. Age categories [0, 18) and [18, 30) were combined because the controls were > 18 years old.

$$logit(p) = \beta^T \cdot I_{(age-category)} \text{where } I \text{ are indicators for each age category.}$$

Under the proposed model (Supplementary Figure S4), a survivor and a control in the same age bin would have a similar chance to develop age-related CH. Therefore, the survivor's probability, q, to develop any CH, which may be age- or therapy-related, was estimated by fixing the age effect as an offset term:

$$logit(q) = logit(\hat{p}) + \theta^T \cdot therapy\ variables\,.$$

The therapy variables included the dose tertiles for alkylating agents, RT, and bleomycin, which were found significantly associated with CH in the multivariable logistic analysis with alkylating agents, RT, bleomycin, platinum, anthracyclines, vinca alkaloids, methotrexate, dactinomycin, epipodophyllotoxins, age at sample collection, and age at diagnosis. The age at diagnosis was also included to capture potential age effects specific to the survivors. Assuming that the age- and therapy-related etiologies were mutually exclusive, CH clones were inferred as therapy-related (iTR) if they belonged to a survivor whose estimated probability for therapy-related CH was higher than the probability for age-related CH:

$$(\hat{q} - \hat{p})/\hat{p} > 1\,.$$

otherwise, inferred as age-related (iAR).

Joint analysis on cell phenotyping and STAT3 Y640F mutation status by Tapestri

The blood samples from three survivors were analyzed; for SJHL018072 a serial sample collected 3 years after the one used for panel sequencing was used due to material exhaustion. Cryopreserved survivor samples were washed with FACS buffer and quantified using a Luna -FL Dual Fluorescence cell counter. Cells ($0.5 – 4.0 \times 10^6$ viable cells) were then resuspended in cell staining buffer (#420201, BioLegend) in the concentration of 25,000 cells/μL and incubated with TruStain FcX, and 1× Tapestri blocking buffer (Mission Bio) for 15 min on ice. Then, TotalSeq™-D Human Heme Oncology Cocktail v1.0 (# 399906, BioLegend), which contains the pool of 45 oligo-conjugated antibodies,

was added and incubated for 30 min on ice. Cells were then washed three times with prechilled cell staining buffer (CSB #420201, BioLegend) followed by resuspension of the cells in the Tapestri Cell suspension buffer (Mission Bio). We followed Tapestri Single-cell DNA+Protein sequencing v2 (custom panel) user guide to process the samples for the single cell amplicon-based DNA and Protein sequencing. In brief, after counting the resuspended cells on an automated cell counter (Luna Biosystems), resuspended cells (2,500 –3,500 cells/μL) were encapsulated using a Tapestri microfluidics cartridge and lysed. A forward primer mix (30 μM each) for the antibody tags was added before barcoding. Barcoded samples were then subjected to targeted PCR amplification of a custom amplicon covering the *STAT3 Y640F* locus. DNA PCR products were then isolated from individual droplets and purified with Ampure XP beads as per the user's guide. The DNA PCR products were then used as a PCR template for library generation with additional four more PCR cycles and repurified using Ampure XP beads. Protein PCR products were incubated with Tapestri pullout oligo (5 μM) at 96 °C for 5 min followed by incubation on ice for 5 min. Protein PCR products were then purified using streptavidin beads (Mission Bio) and were used as a PCR template for the incorporation of i5/i7 Illumina indices followed by purification using Ampure XP beads. All libraries, both DNA and protein, were quantified using an Agilent Bioanalyzer and pooled for sequencing on an Illumina Nextseq 550 or NovaSeq 6000. The resulting FASTQ files for single-cell DNA and protein libraries were uploaded via the Tapestri portal for analysis by Tapestri's pipeline. This pipeline determines the genotype by analyzing the amplified DNA segment and the phenotype by reading the oligo-tag conjugated to the antibody in the same cell that is indexed by i5/i7 Illumina sequences. The analysis result contains the two-dimensional coordinate in the cluster space for each cell as well as meta information for the genotype and phenotype, which was visualized in Fig. 3F. The same data were used to summarize cellular fraction (CF) of *STAT3*-mutant cells in each cell type as shown in Supplementary Figure S5.

Single cell whole genome amplification and STAT3 Y640F genotyping

Thawed survivor blood samples (the same as those used for Tapestri assay), which had been cryopreserved, were washed, and resuspended in culture medium. The resuspended cells were stained with Calcein and DAPI for viable cell sorting. Calcein positive/DAPI negative cells sorted into 96 well DNA loBind Semi Skirted plates containing 4 μL of cell storage buffer (Qiagen # 150370). The whole genome amplification reactions were assembled in a Pre-PCR workstation, which was decontaminated with ultraviolet irradiation before every experiment. Multiple displacement amplification (MDA) was carried out according to the Repli-g Advanced DNA single cell kit (Qiagen # 150365). The amplified WGA yield was in the range of 25–40 μg, stored at −20 °C until further processing *STAT3* genotyping and sequencing library preparation. Aliquots from the amplified DNA material was diluted to 25 ng/μL and the targeted region was amplified by PCR using Q5® Hot Start High-Fidelity 2× Master Mix (New England BioLabs, Cat# M0494L). Two primer sets were used to generate the amplicons. The first set included the forward primer GGAAAGAAAAAATGGGCAG and the reverse AAATCAACAACTACCTGGG. The second set was TTAAGTCTTTTCCCCTTCG for forward and TCAACAACTACCTGGGTC for reverse, respectively. The PCR fragments were cleaned up using ExoSAP-IT PCR product cleanup reagent (ThermoFisher Scientific,

Cat# 78202). The chromatogram from Sanger sequencing was visually inspected for the *STAT3 Y640F* mutation.

Single cell whole genome sequencing (scWGS) and variant identification

For amplified DNA materials from *STAT3 Y640F* or wildtype cells, Illumina compatible whole genomic DNA libraries were constructed using the Kapa HyperPrep kit (Roche, Cat# 07962363001) and IDT for Illumina TruSeq DNA UD Indexes (Illumina, Cat#20023784). Library QC were performed using Bioanalyzer High Sensitivity DNA Analysis chip (Agilent, Cat# 5067-4626) and MiSeq Nano kit (Illumina, Cat# 15036717). The scWGS libraries were denatured with PhiX spike-in, and sequencing was performed on a NovaSeq 6000 using S2 or S4 flow cells and standard workflow to generate 150 cycle paired end reads. The reads were mapped to the GRCh37 reference genome by BWA. The mapped datasets were required to cover >70% of genome at a depth of 10 reads or more.

GATK HaplotypeCaller (4.0.2.1) was applied at the default setting to the single-cell datasets and the bulk-WGS dataset (20) from the same survivor. GATK's VariantFiltration was used to filter the raw calls: --filter-expression "QD < 2.0" --filter-expression "FS > 60.0" --filter-expression "MQ < 40.0" --filter-expression "MQRankSum < −12.5" --filter-expression "ReadPosRankSum < −8.0" --filter-expression "SOR > 3.0" --filter-expression "QUAL < 30" -window 10. MDA reaction tends to yield a non-uniform amplification and, thus, could provide a biased genotype representation due to allele dropout. To select datasets with relatively uniform amplification, we performed a QC check which requires 10x coverage in >70% of the human genome in scWGS, and >90% genotype concordance of heterozygous SNPs in bulk WGS sample in regions with > 10× coverage in both bulk WGS and scWGS. For scWGS data that passed QC, SNVs private to scWGS (i.e., absent in bulk WGS) were considered putative somatic SNVs. Of the scWGS data generated from three HL survivors, only those from SJHL018072 passed the QC check (93.7 – 95.7%), while the other two (SJHL017906 and SJHL019340) failed (59.1 – 87.2%) despite repeated effort for all available DNA materials. From the QC-failed case, SNVs unique to scWGS were also collected for the purpose of building a panel of artifactual SNVs caused by MDA. The final mutation set for SJHL018072 was prepared by filtering with the SNVs recurring in the other two cases, which may represent MDA/sequencing/ artifacts as suggested by the presence of SBS57, an error signature.

Mutational signature analysis

We ran signature profiler (https://github.com/AlexandrovLab/SigProfilerSingleSample) to query against COSMIC signatures (v3.1, GRCh37) with the following options: ref="GRCh37" and exome=False. Given the elevated mutation burden in *STAT3 Y640F* mutant cells compared with the wild-type, we performed the analysis by a) combining all SNVs from the mutant cells to generate a profile for the mutant cells; and b) combining all SNVs from the wild-type cells to generate a profile for the wildtype cells. The signatures from these two categories are shown in Fig. 3G.

Serial sample analysis

Eighty-four survivors had serial blood samples collected from multiple SJLIFE visits. These samples were analyzed for CH variants as described in the preceding section of CH variant detection. For the pair of time points in each serial sample, the growth direction of a CH clone was determined as expanding or non-expanding by comparing the two VAFs with a binomial test or taking the concordance of the ddPCR concentration estimates for samples where the readings were available at both time points.

Simulation study

The stochastic clone dynamics model proposed by Watson *et al.* (33) was used to simulate the age-related CH dynamics. In addition to the original model, a modified model was prepared to study trajectories with early onsets as in CH-positive survivors. This model elevates the mutation rate (original rate by Watson *et al.* (33): 3×10^{-6} /cell) by 5 – 10 times during a random time window starting at age 0 – 15 with a span of 1 – 5 years. Trajectories starting within the hypermutative windows were analyzed. To simulate the serial sample analysis, clones were subjected to a two-timepoint sampling if VAF stayed > 0.1% (the detection limit of this study). The two VAFs were compared to label the clone as expanding or non-expanding.

Computations

Statistical analyses, simulation, and visualization were performed in Python. Comparisons were considered significant if two-sided p-values were < 0.05.

Data Availability

The BAM files for panel sequencing and single cell WGS are available on St. Jude Cloud under accession SJC-DC-1020 (https://platform.stjude.cloud/data/cohorts?dataset_accession=SJC-DS-1020).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENT

We would like to thank Mr. David Rosenfield and team for supporting the analysis of panel sequencing and scWGS data, and Mr. Michael Edmonson for his help on proof-reading the manuscript. We are grateful to Ms. Delaram Rahbarinia, Mr. Michael Macias, and Mr. Clay McLeod for uploading the sequencing datasets to St Jude Cloud. We also thank Ms. Jessica Baedke, Ms. Kyla Shelton, Ms. Huiqi Wang, and Dr. Emily Finch for their effort in collecting the clinical information. We thank Dr. Lois Staudt on discussions related to *STAT3* mutation in Hodgkin lymphoma. Finally, we would like to acknowledge the very insightful comments from the two reviewers, which helped us to substantially improve the quality of this study.

FUNDING

This study was supported by a Cancer Center Support (CORE) Grant (P30 CA21765) to St. Jude Children's Research Hospital, R01 CA216345 (MPIs: Yutaka Yasui and Jinghui Zhang, R01 CA216391 (PI: Jinghui Zhang), U01 CA195547 1 (MPIs: Melissa M. Hudson and Kirsten K. Ness), and the American Lebanese Syrian Associated Charities (ALSAC), Memphis, TN.

REFERENCES

1. Köhnke T, Majeti R. Clonal hematopoiesis: from mechanisms to clinical intervention. Cancer Discov 2021;11:2987–2997. [PubMed: 34407958]
2. Coombs CC, Zehir A, Devlin SM, Kishtagari A, Syed A, Jonsson P, et al. Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. Cell Stem Cell 2017;21:374–382. [PubMed: 28803919]
3. Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nat Genet 2020;52:1219–1226. [PubMed: 33106634]
4. Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC, et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med 2014;20:1472–1478. [PubMed: 25326804]
5. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, Lindsley RC, et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med 2014;371:2488–98. [PubMed: 25426837]
6. Genovese G, Kähler AK, Handsaker RE, Lindberg J, Rose SA, Bakhoum SF, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N Engl J Med 2014;371:2477–87. [PubMed: 25426838]
7. Desai P, Mencia-Trinchant N, Savenkov O, Simon MS, Cheang G, Lee S, et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. Nat Med 2018;24:1015–1023. [PubMed: 29988143]
8. Ness KK, Kirkland JL, Gramatges MM, Wang Z, Kundu M, McCastlain K, et al. Premature physiologic aging as a paradigm for understanding increased risk of adverse health across the lifespan of survivors of childhood cancer. J Clin Oncol 2018;36:2206–2215. [PubMed: 29874132]
9. Robison LL, Hudson MM. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. Nat Rev Cancer 2014;14:61–70. [PubMed: 24304873]
10. Collord G, Park N, Podestà M, Dagnino M, Cilloni D, Jones D, et al. Clonal haematopoiesis is not prevalent in survivors of childhood cancer. Br J Haematol 2018;181:537–539. [PubMed: 28369776]
11. Coorens THH, Collord G, Lu W, Mitchell E, Ijaz J, Roberts T, et al. Clonal hematopoiesis and therapy-related myeloid neoplasms following neuroblastoma treatment. Blood 2021;137:2992–2997. [PubMed: 33598691]
12. Bertrums EJM, Rosendahl Huber AKM, de Kanter JK, Brandsma AM, van Leeuwen AJCN, Verheul M, et al. Elevated mutational age in blood of children treated for cancer contributes to therapy-related myeloid neoplasms. Cancer Discov 2022;12:1860–1872. [PubMed: 35678530]
13. Yeh JM, Ward ZJ, Chaudhry A, Liu Q, Yasui Y, Armstrong GT, et al. Life expectancy of adult survivors of childhood cancer over 3 decades. JAMA Oncol 2020;6:350–357. [PubMed: 31895405]
14. Hudson MM, Ness KK, Gurney JG, Mulrooney DA, Chemaitilly W, Krull KR, et al. Clinical ascertainment of health outcomes among adults treated for childhood cancer. JAMA 2013;309:2371–2381. [PubMed: 23757085]
15. Hudson MM, Ehrhardt MJ, Bhakta N, Baassiri M, Eissa H, Chemaitilly W, et al. Approach for classification and severity grading of long-term and late-onset health events among childhood cancer survivors in the St. Jude Lifetime Cohort. Cancer Epidemiol Biomarkers Prev 2017;26:666–674. [PubMed: 28035022]
16. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol 2019;20:50. doi: 10.1186/s13059-019-1659-6. [PubMed: 30867008]
17. Liu FT, Ting KM, Zhou ZH. Isolation forest. In 2008 eighth IEEE international conference on data mining 2008 Dec 15 (pp. 413–422). IEEE.
18. Hagiwara K, Edmonson MN, Wheeler DA, Zhang J. indelPost: harmonizing ambiguities in simple and complex indel alignments. Bioinformatics 2022;38:549–551. [PubMed: 34431982]
19. Song N, Li Z, Qin N, Howell CR, Wilson CL, Easton J, et al. Shortened leukocyte telomere length associates with an increased prevalence of chronic health conditions among survivors of childhood

cancer: a report from the St. Jude Lifetime cohort. Clin Cancer Res 2020;26:2362–2371. [PubMed: 31969337]

20. Wang Z, Wilson CL, Easton J, Thrasher A, Mulder H, Liu Q, et al. Genetic risk for subsequent neoplasms among long-term survivors of childhood cancer. J Clin Oncol 2018;36:2078–2087. [PubMed: 29847298]
21. VanderWeele TJ. Mediation Analysis: A Practitioner's Guide. Annu Rev Public Health. 2016;37:17–32. [PubMed: 26653405]
22. Qin N, Wang Z, Liu Q, Song N, Wilson CL, Ehrhardt MJ, et al. Pathogenic germline mutations in DNA repair genes in combination with cancer treatment exposures and risk of subsequent neoplasms among long-term survivors of childhood cancer. J Clin Oncol 2020;38:2728–2740. [PubMed: 32496904]
23. Pich O, Reyes-Salazar I, Gonzalez-Perez A, Lopez-Bigas N. Discovering the drivers of clonal hematopoiesis. Nat Commun 2022;13(1):4267. [PubMed: 35871184]
24. Koskela HL, Eldfors S, Ellonen P, van Adrichem AJ, Kuusanmäki H, Andersson EI, et al. Somatic STAT3 mutations in large granular lymphocytic leukemia. N Engl J Med 2012;366:1905–13. [PubMed: 22591296]
25. Küppers R, Rajewsky K, Zhao M, Simons G, Laumann R, Fischer R, et al. Hodgkin disease: Hodgkin and Reed-Sternberg cells picked from histological sections show clonal immunoglobulin gene rearrangements and appear to be derived from B cells at various stages of development. Proc Natl Acad Sci U S A 1994;91:10962–6. [PubMed: 7971992]
26. Marafioti T, Hummel M, Anagnostopoulos I, Foss HD, Falini B, Delsol G, et al. Origin of nodular lymphocyte-predominant Hodgkin's disease from a clonal expansion of highly mutated germinal-center B cells. N Engl J Med 1997;337:453–8. [PubMed: 9250847]
27. Machado HE, Mitchell E, Øbro NF, Kübler K, Davies M, Leongamornlert D, et al. Diverse mutational landscapes in human lymphocytes. Nature 2022;608:724–732. [PubMed: 35948631]
28. Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. Cell 2019;176:1282–1294. [PubMed: 30849372]
29. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 2019;574:532–537. [PubMed: 31645730]
30. Santarsieri A, Mitchell E, Sturgess K, Brice P, Menne TF, Osborne W, et al. Replacing procarbazine with dacarbazine in escalated BEACOPP dramatically reduces the post treatment haematopoietic stem and progenitor cell mutational burden in Hodgkin lymphoma patients with no apparent loss of clinical efficacy. Blood 2022;140(Supplement 1):1761–4.
31. Acuna-Hidalgo R, Sengul H, Steehouwer M, van de Vorst M, Vermeulen SH, Kiemeney LALM, et al. Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. Am J Hum Genet 2017;101:50–64. [PubMed: 28669404]
32. Mitchell E, Spencer Chapman M, Williams N, Dawson KJ, Mende N, Calderbank EF, et al. Clonal dynamics of haematopoiesis across the human lifespan. Nature 2022;606:343–350 [PubMed: 35650442]
33. Watson CJ, Papula AL, Poon GYP, Wong WH, Young AL, Druley TE, et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science 2020;367:1449–1454. [PubMed: 32217721]
34. Galluzzi L, Buqué A, Kepp O, Zitvogel L, Kroemer G. Immunological effects of conventional chemotherapy and targeted anticancer agents. Cancer Cell 2015;28:690–714. [PubMed: 26678337]
35. Caiado F, Pietras EM, Manz MG. Inflammation as a regulator of hematopoietic stem cell function in disease, aging, and clonal selection. J Exp Med 2021;218:e20201541. [PubMed: 34129016]
36. Laconi E, Marongiu F, DeGregori J. Cancer as a disease of old age: changing mutational and microenvironmental landscapes. Br J Cancer 2020;122:943–952. [PubMed: 32042067]
37. Howell CR, Bjornard KL, Ness KK, Alberts N, Armstrong GT, Bhakta N, et al. Cohort Profile: The St. Jude Lifetime Cohort Study (SJLIFE) for paediatric cancer survivors. Int J Epidemiol 2021 Mar;50:39–49. [PubMed: 33374007]

38. Young AL, Challen GA, Birmann BM, Druley TE. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat Commun 2016;7:12484. [PubMed: 27546487]

39. Edmonson MN, Zhang J, Yan C, Finney RP, Meerzaman DM, Buetow KH. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. Bioinformatics 2011;27:865–866. [PubMed: 21278191]

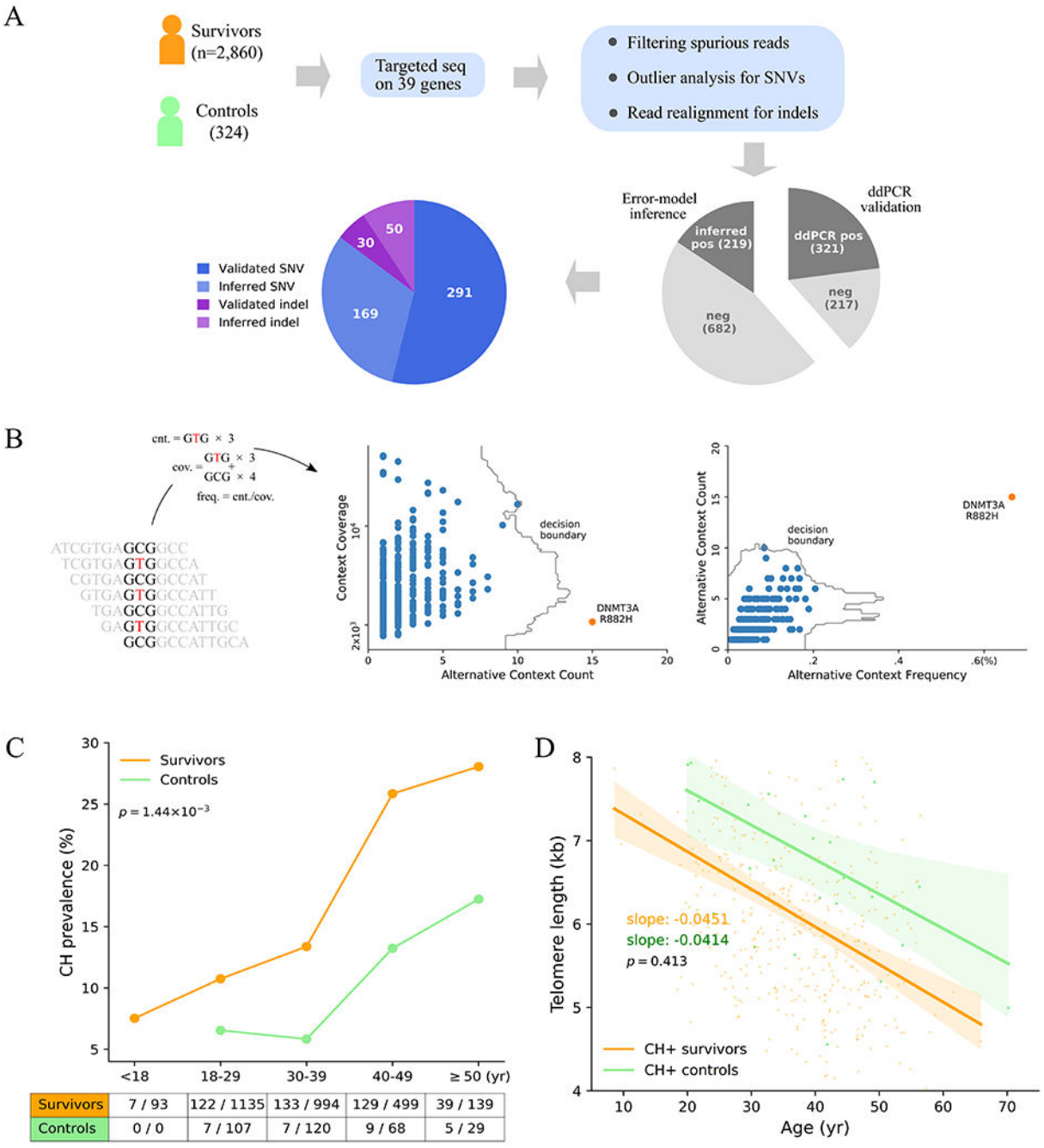
40. Hagiwara K, Ding L, Edmonson MN, Rice SV, Newman S, Easton J, et al. RNAIndel: discovering somatic coding indels from tumor RNA-Seq data. Bioinformatics 2020;36:1382–1390. [PubMed: 31593214]

41. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, et al. Germline mutations in predisposition genes in pediatric cancer. N Engl J Med 2015;373:2336–2346. [PubMed: 26580448]

42. Cristy M. Active bone marrow distribution as a function of age in humans. Phys Med Biol 1981;26:389–400. [PubMed: 7243876]

STATEMENT OF SIGNIFICANCE

This first comprehensive CH analysis in long-term survivors of pediatric cancer presents the elevated prevalence and therapy-exposures/diagnostic spectrum associated with CH. Due to the contrasting dynamics of clonal expansion for age-related versus therapy-related CH, longitudinal monitoring is recommended to ascertain the long-term effects of therapy-induced CH in pediatric cancer survivors.



	<18	18-29	30-39	40-49	≥ 50 (yr)
Survivors	7 / 93	122 / 1135	133 / 994	129 / 499	39 / 139
Controls	0 / 0	7 / 107	7 / 120	9 / 68	5 / 29

Figure 1. Elevated CH in pediatric cancer survivors.

A, Study design for CH analysis in survivors. Blood samples from survivors and controls analyzed by capture sequencing on 39 CH-associated genes. Putative CH variant analysis involves computational error suppression of sequencing reads, outlier analysis of substitutions with the same genomic context, and indel re-alignment. Approximately 40% of the putative variants (538 total) were successfully assayed by ddPCR, while the validity of the remaining variants was inferred by comparing to a background error model. **B,** Outlier analysis based on the sequence context of a CH variant. For each of the 96 genomic triplet

context changes, the count and frequency for sequencing context matching the context of the alternative allele as well as the read count were within a sample. Outlier analysis was performed by IsolationForest (17) in both the spaces spanned by the count and coverage, and count and frequency. Example shows the GCG>GTC context analysis in an acute lymphoblastic leukemia survivor, in which the *DNMT3A R882H* mutation was detected as an outlier. **C,** Elevated CH prevalence in the survivors (orange) compared to the controls (green) across all five age categories. Significance of difference in the overall prevalence of the two groups was based on Fisher's exact test. **D,** Leukocyte telomere lengths in the CH positive survivors and controls. Telomere attrition rates were estimated as a regression slope, which was compared by t-test. Bands represent 95% CIs of the least-square regressions.

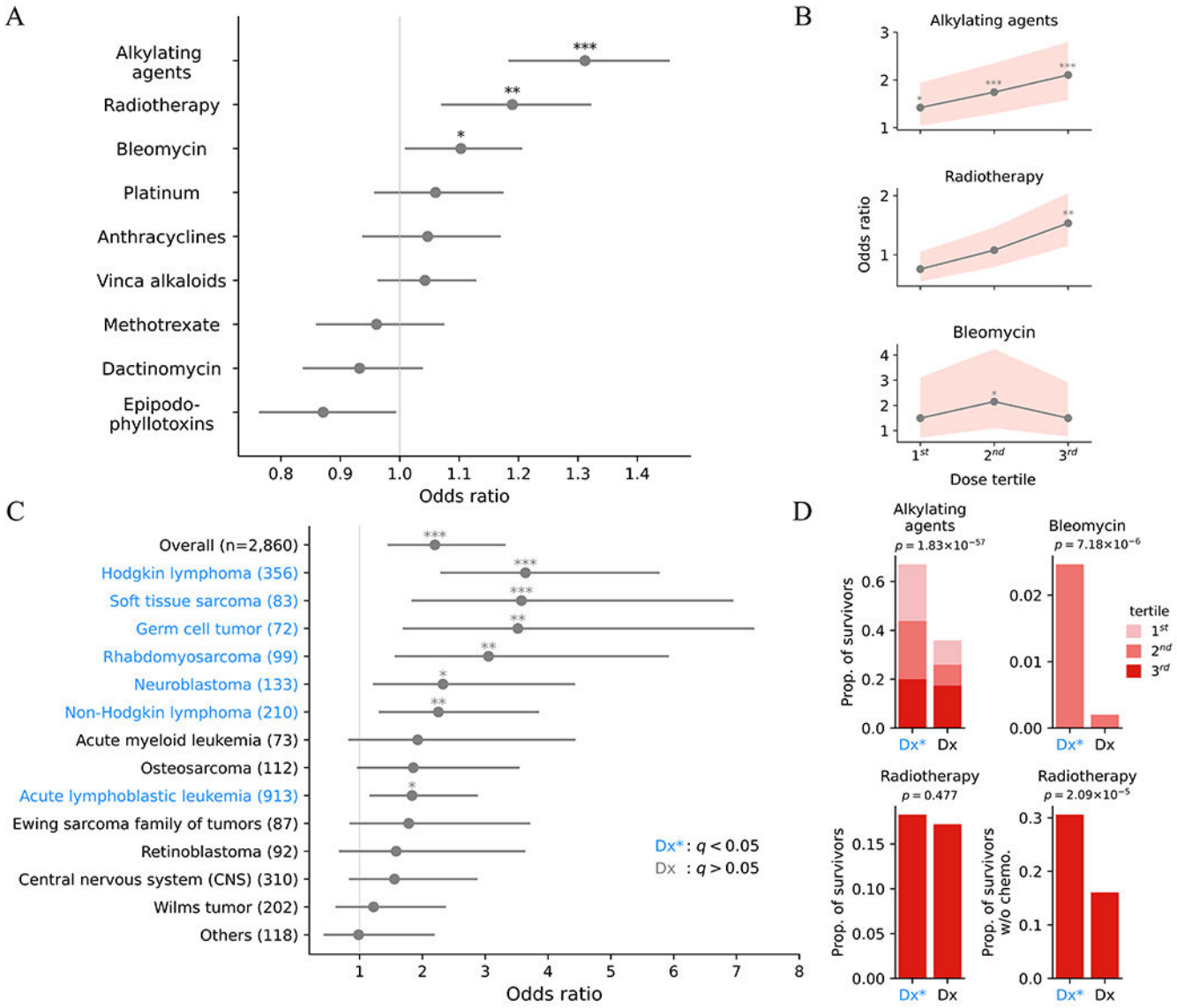


Figure 2. Associations of CH prevalence with clinical variables.

A, Odds ratios and 95% CIs for CH and cancer therapies. Multivariable logistic analysis adjusted for each other therapy doses, age at blood sampling, and age at diagnosis. Cumulative doses were used for chemotherapy. For RT, maximum region-specific dose was used after adjusting for the active bone marrow distribution at the diagnosis age. Variables were scaled by standard deviation. Bars represent 95% CIs. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. **B,** Odds ratio between significant therapies in **A** and CH over the dose tertile. Odds ratios were similarly adjusted as in **A**. Bands represent 95% CIs. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. **C,** Age-adjusted odds ratios of CH relative to the control were estimated for diagnoses with 50 survivors or more. Uncommon diagnoses with < 50 cases were combined as “Others”. Diagnoses in blue indicate a significant odds ratio and are collectively denoted as *Dx**. Bars represent 95% CIs. *q (Benjamini-Hochberg corrected p) < 0.05, **$q < 0.01$, ***$q < 0.001$. **D,** Comparison of therapy intensity in CH-associated cancer types (*Dx**,

blue letter) with those not associated with CH (*Dx*, black letter). The frequency of therapy intensities found significant in **B** was compared between the two groups by proportion test. The analysis on the bottom right was restricted to survivors who had not received chemotherapy.

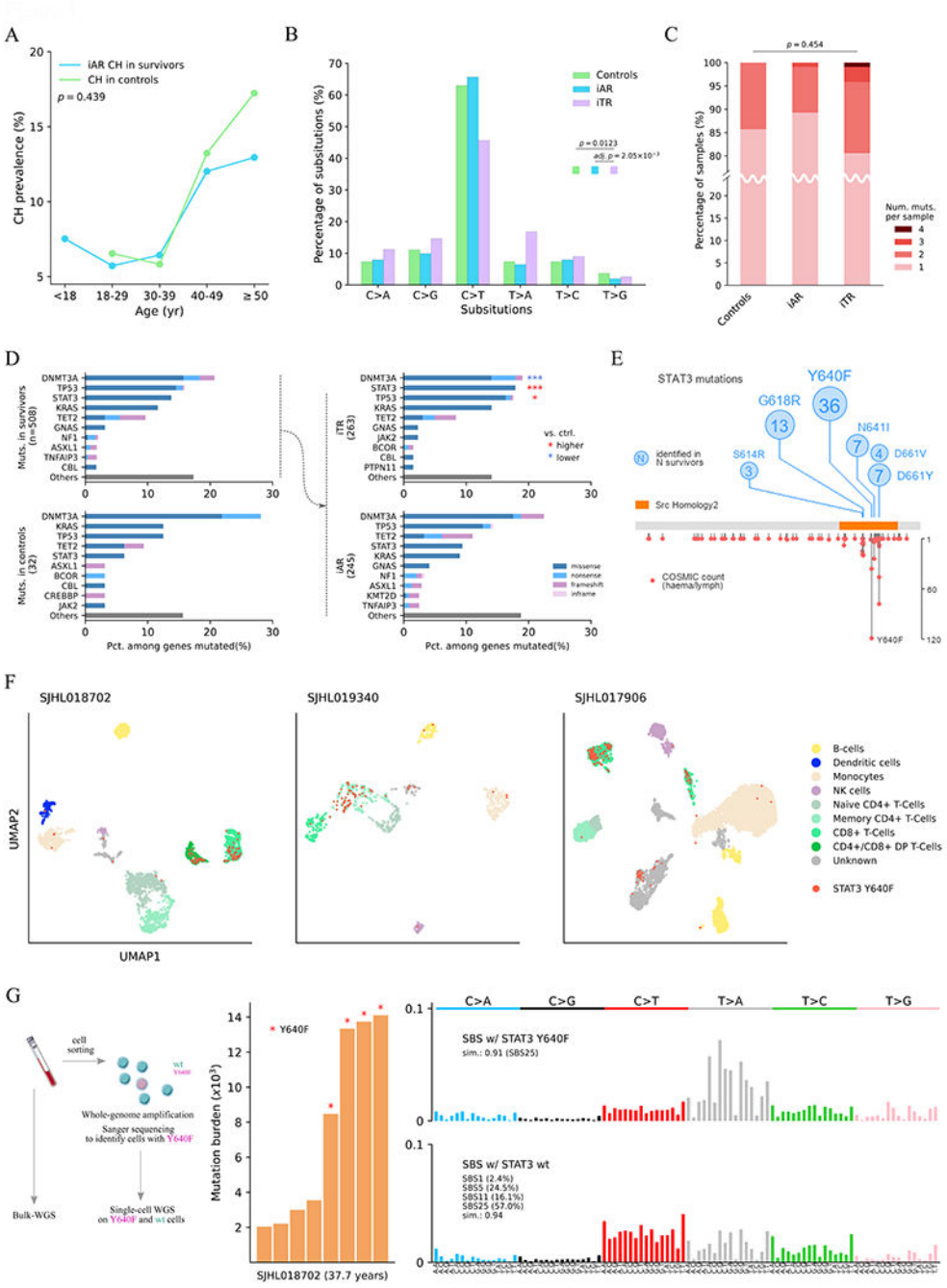


Figure 3. Molecular features of inferred age- versus therapy-related CH.

A, Comparable prevalence of inferred age-related (iAR) CH in the survivors (blue) and CH in the controls (green). Fisher's exact test showed no significant difference between the two groups ($p = 0.439$). **B**, Mutation spectra of SNVs in the control (green), iAR (blue) and iTR (purple). Spectrum frequency was compared by χ^2 test followed by post hoc χ^2. Only significant comparisons are shown. Post hoc p-value was adjusted by Šidák's method. **C**, Distribution of mutation count per sample. The distribution in each group was compared by Kruskal–Wallis test. **D**, CH mutation frequency in the top 10 genes identified

in the survivors and controls. The CH mutations in survivors were further stratified into iAR and iTR groups and the top 5 genes were compared to the control by two-sided proportion test. *q (Benjamini-Hochberg corrected p) < 0.05 and ***q < 0.001 with red indicating a proportion higher than control, and blue if lower. **E**, *STAT3* mutations identified in this study (*top*, number in the circle represents occurrence) and the COSMIC (v97) for hematological and lymphoid tumors (*bottom*, y-axis shown occurrence). **F**, Cell phenotype and *STAT3 Y640F* genotype in blood samples from three survivors profiled by Mission Bio's Tapestri assay on single cell amplicon-based DNA and protein sequencing. Cells were mapped by uniform manifold approximation and projection (UMAP) by protein antibody data and colored by their corresponding cell types. Those harboring *Y640F* mutation are labeled in red. **G**, Mutation burden and signature analysis for SJHL018702 by single-cell whole-genome sequencing (scWGS). The peripheral blood sample was sorted for viable cells followed by multiple displacement amplification (MDA) to generate scWGS data in *STAT3 Y640F*-mutant and wildtype cells (*left*). Mutation burden by somatic SNVs identified from scWGS using bulk WGS as a control is shown as bar plot with *Y640F* mutant cells marked by a red star (*middle*). Mutational signatures of *Y640F* mutant cells and wildtype cells are shown at the right.

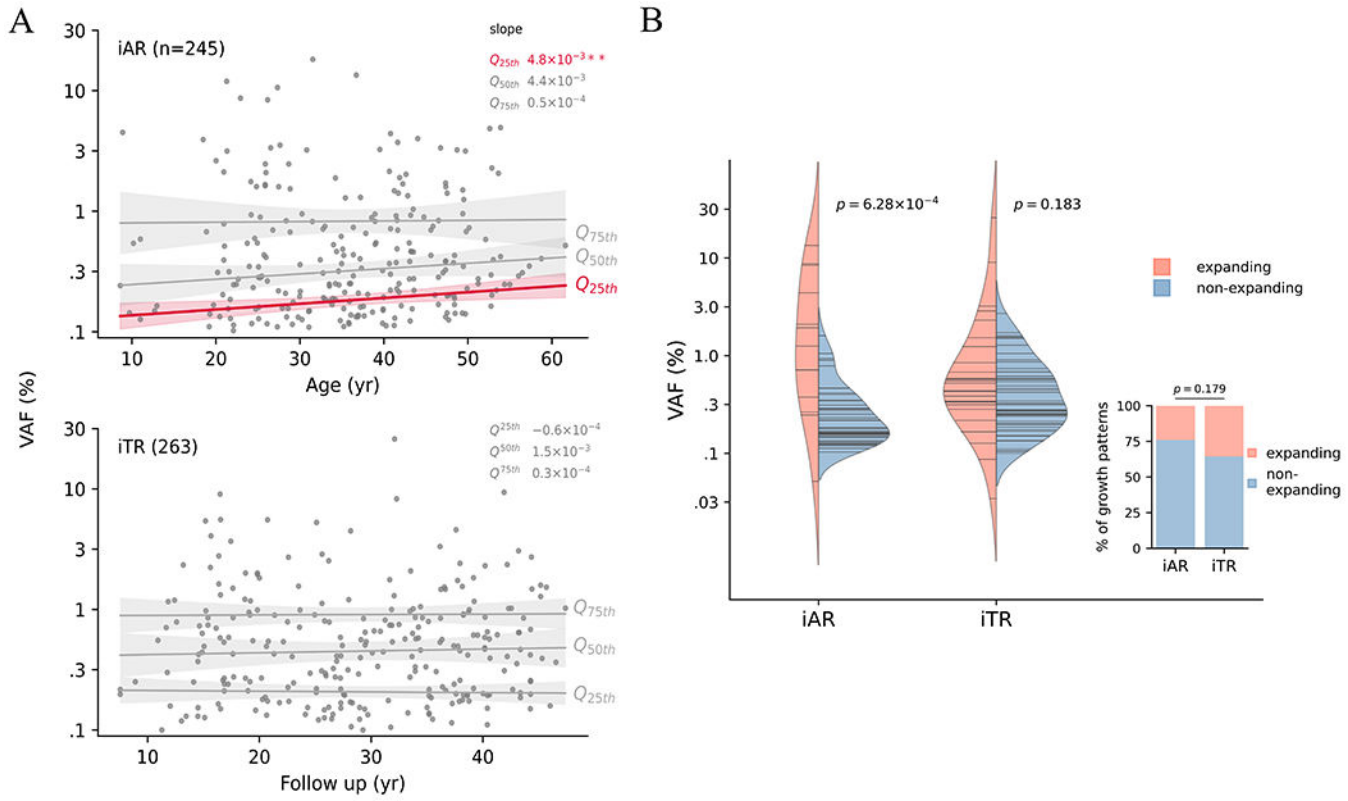


Figure 4. Dynamics of inferred age- versus therapy-related CH.

A, VAF distribution along age for iAR and along follow-up for iTR. Longitudinal trend was estimated by quantile regression on VAF quartiles: Q_{25th} (lower 25%), Q_{50th} (median) and Q_{75th} (upper 25%). Estimated slopes are shown. ** $p < 0.005$. **B**, Association between clone size and subsequent expansion revealed by serial analysis. For each serial sample pair, clone fate was classified as expanding, or non-expanding based on statistical test. The proportion of the expanding fate did not differ between iAR and iTR (**inset**, Fisher exact p=0.179). Log-transformed VAFs at the earlier time point was compared by t-test between expanding and non-expanding fates. Horizontal lines in the violin represent the VAF (log-scale) observed at the initial timepoint.