**SUPPLEMENTARY MATERIALS**

Hagiwara *et al*. Dynamics of age- versus therapy-related clonal hematopoiesis in long-term survivors of pediatric cancer.

Supplementary Methods

1. Survivor cohort2

2. Targeted deep sequencing2

3. Variant detection3

4. Orthogonal validation by digital droplet PCR 7

5. Inferring the status of undetermined variants 9

6. Amplicon sequencing for *PPM1D* 9

7. Schematic representation of the variant detection workflow 11

Supplementary Figures

Figure S112

Figure S213

Figure S314

Figure S415

Figure S516

Figure S617

Figure S718

Figure S819

Figure S920

Figure S1021

Supplementary Tables

Table S122

Table S223

Table S324

Table S425

Table S526

Table S627

Table S7-928

Supplementary References29

**Supplementary Methods**

**1. Survivor cohort**

We characterized CH of 2,860 survivors and 324 community controls who participated in the St. Jude Lifetime Cohort (SJLIFE) study. All survivors were previously analyzed by Wang *et al.* (1) for whole-genome sequencing (WGS) and whole-exome sequencing (WES) as part of the 3,006 survivors-341 control cohort. Of the 3,006 survivors, this study excluded those with insufficient DNA materials (n = 97) and survivors whose available materials were collected < 5 years since diagnosis (n = 12). Of the 2,897 survivors subjected to targeted deep sequencing, 37 samples failed to pass the quality check due to low coverage (see **2.** **Targeted deep sequencing**). Our study only included survivors who were alive at SLIFE enrollment. To evaluate the potential bias due to loss of survivors with secondary leukemia, we reviewed the full survivor cohort at St. Jude, which consists of 9,358 cases achieving 5-year survival. Secondary leukemias/MDS developed in 105 survivors, 64 of whom died prior to recruitment into the SJLIFE program. This represented <1% (64/9,358) of all survivors who were treated in our institution.

**2. Targeted deep sequencing**

Illumina compatible genomic DNA libraries were constructed using either the Illumina Nextera DNA library prep kit, or the Kapa hyper DNA library prep kit. The libraries were hybridized to an Integrated DNA Technologies xGEN lockdown mini pool probe mix covering coding regions of the following genes:  *APC, ASXL1, BCOR, BRCA1, BRCA2, BRCC3, CBL, CREBBP, CUX1, DICER1, DNMT3A, GNAS, GNB1, IDH2, JAK2, KMT2D, KRAS, LUC7L2, MSH6, NF1, NRAS, PALB2, PMS2, PRDM1, PTCH1, PTPN11, RAD21, RB1, SDHA, SF3B1, SRSF, STAT3, SUFU, TET2, TNFAIP3, TP53, U2AF1, WT1, ZRSR2*. The hybridization blocking and wash conditions were performed according to manufacturer’s recommendations using 500 ng per library and 16 libraries per hybridization pool. Sequencing was performed on a NovaSeq 6000 using S4 flow cells and the XP workflow with 96 samples per lane to generate 150 cycle paired end reads. The reads were mapped to the human GRCh37 genome by BWA (2) to generate a BAM file (i.e., the compressed binary version of a SAM file that is used to represent aligned sequence) for each sample. Using Phred quality score ≥ 20 as a threshold for high-quality bases, BAM files with ≥ 80% of the targeted region covered with ≥1,000× high-quality bases were qualified for subsequent analysis.

**3. Variant detection**

Somatic variants were separately identified for single nucleotide variants (SNVs) and small insertions/deletions (indels). The workflow is detailed in the following subsections along and a schematic representation is shown in Supplementary Method section 7 “**Schematic representation of the variant detection workflow”** of this document.

**3.1. Single nucleotide variants**

*Outlier detection*:

Each of 96 genomic-triplet substitutions is known to have distinct sequencing error patterns (3). Thus, putative somatic SNVs were detected as those with high read count or allele fraction outlying the background count or fraction based on triplet context. We used CleanDeepSeq (3) which generates base counts at each locus after filtering spurious reads. At each targeted locus, the most abundant non-reference base was considered a putative substitution. To the putative substitution, the immediate 3´ and 5´ reference bases were added to construct the alternative triplet context. The sum of the context counts was used as the context coverage and the alternative context frequency was defined as the ratio between alternative context count and context coverage. Substitutions with alternative context frequency ≥ 0.35 were excluded because they may be germline events. The vast majority of low frequency variants likely arise as artifacts at high sequencing depth rather than true genomic alterations. Hence, we generated the context-specific error distribution by plotting the counts and frequencies for 96 pairs of reference and alternative contexts within a sample. Outliers of alternative context count vs. context coverage distribution or alternative context count vs. alternative context frequency distribution, detected by Isolation Forest (4), were considered candidate somatic mutations.

*Variant selection and filtering:*

To select variants relevant for CH analysis, we retained SNVs annotated as missense/nonsense (5) if they occurred a minimum of 5 times in the COSMIC (v90) (6) database. For genes known to be recurrently mutated in clonal hematopoiesis, i.e., *ASXL1*, *DNMT3A*, or *TET2* (7 ‒ 10), SNVs in the following highly mutated regions were retained if VAF > 0.005 regardless of their COSMIC count: nonsense SNVs in *ASXL1* exon 12‒13 (NM015338); nonsynonymous SNVs in *DNMT3A* exon 7‒23(NM175629); missense SNVs in *TET2* exon 4‒10 or aa1853‒1905 in exon 11 (NM001127208), and any nonsense SNVs. For the remaining variants, we excluded those with high recurrence within the St Jude Life cohort but low incidence (n < 5) in COSMIC. The exclusion criteria as follows: i) present in homopolymer repeats > 3; ii) absent in COSMIC but detected in ≥ 3 SJLIFE samples; iii) detected once in COSMIC but in ≥ 5 SJLIFE samples; iv) detected 2 to 4 times in COSMIC but in ≥ 10 SJLIFE samples; and v) Variants with the non-cancer population allele frequency in gnomAD (r2.1.1) ≥ 0.001 were filtered.

*Read count and VAF filtering:*

Outliers were required to have a minimum of alternative context count ≥ 10. Three additional exclusion filters were applied: i) variant allele is supported only by reads from one orientation; ii) variant allele fraction (VAF) <0.001; iii) raw VAF calculated from all sequencing reads was > 50% lower than the CleanDeepSeq-based alternative context frequency—we implemented this filter because high discrepancy may reflect substantial technical error at the locus.

**3.2. Indels**

*Coding indel detection*:

Small insertion/deletions (indels) were detected by Bambino (11) modified for indel detection (12). The indel calls were post-processed by indelPost (13) for allele count correction and complex indel annotation. Indels annotated (5) in coding region were kept for further analysis.

*Variant selection and filtering:*

Indels were filtered if one of the following criteria was met; i) the non-cancer population allele frequency in gnomAD ≥ 0.001; ii) absent in COSMIC but detected in ≥ 3 SJLIFE samples; iii) detected in ≥ 10 SJLIFE samples.

*Read count and VAF filtering:*

Owing to the diminished hybridization efficiency for indel sequences, we set the threshold for excluding germline indel at a VAF of 0.2 (instead of the 0.35 of germline SNVs). Putative somatic indels were kept if all of the following criteria were satisfied: i) VAF (0.001, 0.2); ii) Variant allele detected bidirectionally; iii) indels in repeat were required to have 30 supporting reads or more.

*Manual review:*

Each indels were visually inspected for the following exclusion criteria: i) occurred in a cluster of mutations or region with low-quality calls; ii) only supported in read ends; iii) occurred in low-complex region such as non-exact repeats and did not pass the criteria applied for repeat indels; iv) germline nature was suspected by checking the accompanying WGS data and the presence in dbSNP database; v) realignment anomalies such as over allele counting were suspected; vi) indel reads were not uniquely remapped by BLAT (14).

**3.3. Performance evaluation**

We benchmarked the outlier detection pipeline developed for this project (described in **3.1.**) against MuTect2 (v4.1.8.0), which supports deep-sequencing analysis for variants present at very low variant allele fraction (VAF). Somatic mutations for benchmarking (n = 18, detailed in table below) were prepared by using a matched cancer/normal cell line COLO829/COLO829BL (3). Due to the polyploidy, diluting the cancer genomic DNA of COLO829 with the control DNA of COLO829BL results in four distinct VAF levels. For benchmark analysis, we used a deep-sequencing dataset (SRR7251197) targeting over the loci ± 100bp of region of the benchmarking mutations at a 1:1000 dilution, which contains the somatic mutations at 0.05%, 0.1%, 0.13%, and 0.2% of VAF. We down-sampled reads to approximate the coverage to the current study (13,901× after downsampling) and performed variant detection using MuTect2 and our approach described in 3.1. MuTect2 was applied with the following options: --initial-tumor-lod 0 --tumor-lod-to-emit 0 --max-reads-per-alignment-start 0 --pruning-lod-threshold -4. Our pipeline was applied as described under *Outlier detection* in **3.1.** As shown in the table below, MuTect2 and our method detected 2 and 13 of the benchmarking mutations, respectively. The number of false positives, which were defined as other calls from the target region, was 6 for MuTect2 and 2 for the current approach.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CHROM | POS (hg19) | REF | ALT | Ploidy a | | VAF | | MuTect2 | | Current Method | |
| chr1 | 150208310 | C | G | 1/4 | 0.05 | |  | |  | |
| chr1 | 153391771 | G | T | 2/4 | 0.1 | | ✓ | | ✓ | |
| chr1 | 154002369 | T | C | 4/4 | 0.2 | |  | | ✓ | |
| chr1 | 154128614 | A | C | 2/4 | 0.1 | |  | |  | |
| chr1 | 155264487 | C | T | 2/4 | 0.1 | |  | |  | |
| chr1 | 155300719 | C | T | 2/4 | 0.1 | |  | | ✓ | |
| chr1 | 166991079 | C | G | 2/4 | 0.1 | |  | | ✓ | |
| chr1 | 168105565 | G | A | 2/4 | 0.1 | |  | | ✓ | |
| chr1 | 177251423 | A | G | 1/4 | 0.05 | |  | | ✓ | |
| chr1 | 179604984 | T | C | 4/4 | 0.2 | |  | | ✓ | |
| chr1 | 196928281 | C | T | 4/4 | 0.2 | |  | | ✓ | |
| chr1 | 197169114 | A | C | 1/4 | 0.05 | |  | | ✓ | |
| chr1 | 206730949 | G | A | 4/4 | 0.2 | | ✓ | | ✓ | |
| chr1 | 210329076 | C | T | 4/4 | 0.2 | |  | | ✓ | |
| chr1 | 231858494 | C | T | 4/4 | 0.2 | |  | | ✓ | |
| chr4 | 4419363 | T | C | 1/2 | 0.1 | |  | |  | |
| chr4 | 48501625 | G | A | 1/2 | 0.1 | |  | |  | |
| chr7 | 140453136 | A | T | 4/6 | 0.13 | |  | | ✓ | |

a The number of mutant alleles (n) over the number of chromosome copies (m) is expressed as n/m. For example, 1/4 denotes one mutant allele in tetraploid. Diluting this mutation into the matched normal DNA at 1:1000 results in 0.05% VAF with one part of mutant allele in a thousand parts of diploid genome (1/2000).

**4****. Orthogonal validation by digital droplet PCR (ddPCR)**

A subset of the CH candidate variants was orthogonally validated by digital droplet PCR (ddPCR). Validation was prioritized for variants in the following genes known to be involved in clonal hematopoiesis and myeloid malignancies: *ASXL1*, *BCOR*, *CBL*, *DNMT3A*, *GNAS*, *GNB1*, *IDH2*, *JAK2*, *KRAS*, *NF1*, *NRAS*, *RAD21*, *SF3B1*, *SRSF2*, *TP53,* and *WT1*. We also validated all *STAT3* variants as enrichment of the *STAT3* variants in the Hodgkin lymphoma survivors is a novel finding in our study. Experimental details of the ddPCR validation were as described below.

Genomic DNA for these assays was extracted from peripheral blood mononuclear cells using the Qiagen blood and tissue kit. Batch primer design to the targets of interest was completed using primerTK (<https://github.com/stjude/PrimerTK>). Primer length was kept between 20‒27 bases, melting temperature (Tm) was set between 54‒58 ºC, with a product size of 70‒130 bases. Locked nucleic acid (LNA) probes for variant detection were synthesized by Integrated DNA Technologies (IDT). Probe lengths were between 10‒15 bases with a Tm 4‒7 ºC above the Tm of the primer pair. Probes targeting the reference sequence were conjugated with the HEX or SUN fluorophore. Probes designed to hybridize to the variant sequence were conjugated with FAM. Alternately, premixed ddPCR assays (primers and probes) were acquired from BioRad Laboratories. Detailed information about the primer and probe sequences can be found in Supplementary Table S6.

A No Template Control (NTC) was included to monitor potential contamination. DNA supercontrol plasmids (IDT) containing on average 10 of the reference or alternate amplicon sequences were used to verify probe specificity for the reference or variant sequence and check signal intensity. A unique restriction enzyme site in the poly linker was used to linearize the plasmid DNA. The ddPCR reaction were performed using 150 ng of genomic DNA, plasmid supercontrols at estimated concentration of 3000 copies/µL, 250 nM of each probe, 450 nM of each primer (900 nM BioRad assays) and 1× ddPCR Supermix for Probes without dUTP (BioRad). Digestion of genomic DNA was done to facilitate Poisson genomic DNA distribution in droplets. Restriction enzymes were chosen such that they did not digest within the amplicon. Restriction enzymes used included MseI, HaeIII, CviQi, ScaI, or ClaI (New England BioLabs Inc). Droplet generation was performed on the BioRad AutoDG QX200 ddPCR platform with the PCR in droplet amplification carried out using a BioRad C1000 thermal cycler as follows: enzyme activation at 95 ºC for 10 min; 40 cycles of denaturation at 94 ºC for 30 sec, annealing/extension 52‒58 ºC for 1 min; enzyme deactivation at 98 ºC for 10 min; 4 ºC hold. Ramp rate was 2 ºC/min. Data analysis was completed using QuantaSoft Analysis Pro software (BioRad).

**5. Inferring the status of untested variants**

The status of the untested variants, i.e., those not selected for ddPCR assay or failed in ddPCR assay, was inferred by estimating the probability that the variant was caused by error. For each untested variant, the background error rate was defined as the number of reads supporting the alternative allele divided by the coverage in 100 randomly selected BAM files from the cohort that excluded the sample of interest. For each individual harboring the variant, a binomial *p*-value was calculated. The status of the ddPCR-assayed variants was used to determine the *p*-value threshold by receiver-operating characteristic analysis. The selected thresholds were *p* = 0.0007 for SNV and 9.46 × 10-10 for indels.

**6. Amplicon sequencing for *PPM1D***

As somatic mutations of *PPM1D* are predominantly located in exons 5 and 6, these two exons were amplified by PCR using the primers listed below (IDT) which included an Illumina compatible 16S NGS annealing sequence. The PCR conditions described in the table below were used in a multiplex reaction to generate the amplicons. After plate preparation, the reactions were amplified in a BioRad C1000 thermocycler using the following cycling conditions: activation 3 minutes 95°C, 12‒18 cycles of amplification consisting of denaturation 98°C for 20 seconds, annealing 61°C for 15 seconds, and extension 72°C for 15 seconds followed by a final extension of 30 second at 72°C. Amplified PCR products were purified using a 0.8×ratio of Ampure XP beads (Beckman Coulter) following the vendor recommendations. Amplicon yield and representation was then determined by loading an aliquot of the sample onto an Agilent Tapestation 4150 using a D1000 screen tape. After verification of the amplification, 20μl of the cleaned product from each sample were used as input into the Illumina 16S NGS protocol with IDT® for Illumina® DNA/RNA UD Indexes Sets A‒D (tagmentation) according to the manufacturer’s recommendations with the following amplification conditions: activation 95°C for 3 minutes, 6 cycles of denaturation at 95°C for 30 seconds, annealing at 62°C for 30 seconds, and extension at 72°C for 30 seconds, followed by a final extension at 72°C for 1 minute. The library products were purified and quantitated as described for the initial amplification reaction. After quantitation the libraries were combined into pools of 384 samples and sequenced on an Illumina MiSeq using a 500 cycle V2 reagent kit to generate 250bp paired end reads. After sequencing the run files were converted to FASTQ files and mapped to GRCh37 using BWA. The resulting data was then examined for the presence of the *PPM1D* variants. Similar to the SNVs identified in the 39 genes, we focused on non-synonymous mutations with COSMIC count ≥ 5, which include nonsense mutations: *W427\**, *R458\**, *C478\**, *R552\** and *R572\**. The loci supporting these variations were genotyped in each sample to record read counts. For each nonsense amino acid change, a binomial error model was constructed using read counts with variants with VAF ≥ 0.1%. Variants with *p*-value < 0.05 after Benjamini–Hochberg procedure were considered positive. Indels were identified as described above for the 39 genes.

|  |  |
| --- | --- |
| Primer | Sequence |
| exon5 fw | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGAGCAGATAACACTAGTGC |
| exon5 rv | GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGACGCTAACCAAAGAACT |
| exon6-i fw | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGCCTAATATCACATGCATAG |
| exon6-i rv | GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAGGCTATTATTCAAAGAATCATG |
| exon6-ii fw | TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGTGCCTACTAATTCAACAAA |
| exon6-ii rv | GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAAACCTCATTTCCCAGATG |

Primer design information

|  |  |  |  |
| --- | --- | --- | --- |
| Item | Quantity per sample (µL) | | Quantity per plate (96 samples) (µL) |
| Genomic DNA (5ng/µl in TE buffer pH 8.5) | 20 | | 20 × 96 |
| Primer pair exon5 (5µM) | 5 | 500nM | 480 |
| Primer pair exon6-i (7.5µM) | 750nM |
| Primer pair exon 6-ii (5µM) | 500nM |
| 2× KAPA HiFi HotStart Polymerase (Roche) | 25 | | 2400 |
| Microseal ‘A’ film (BioRad) | - | | 1 |

PCR condition

**7. Schematic representation of the variant detection workflow**

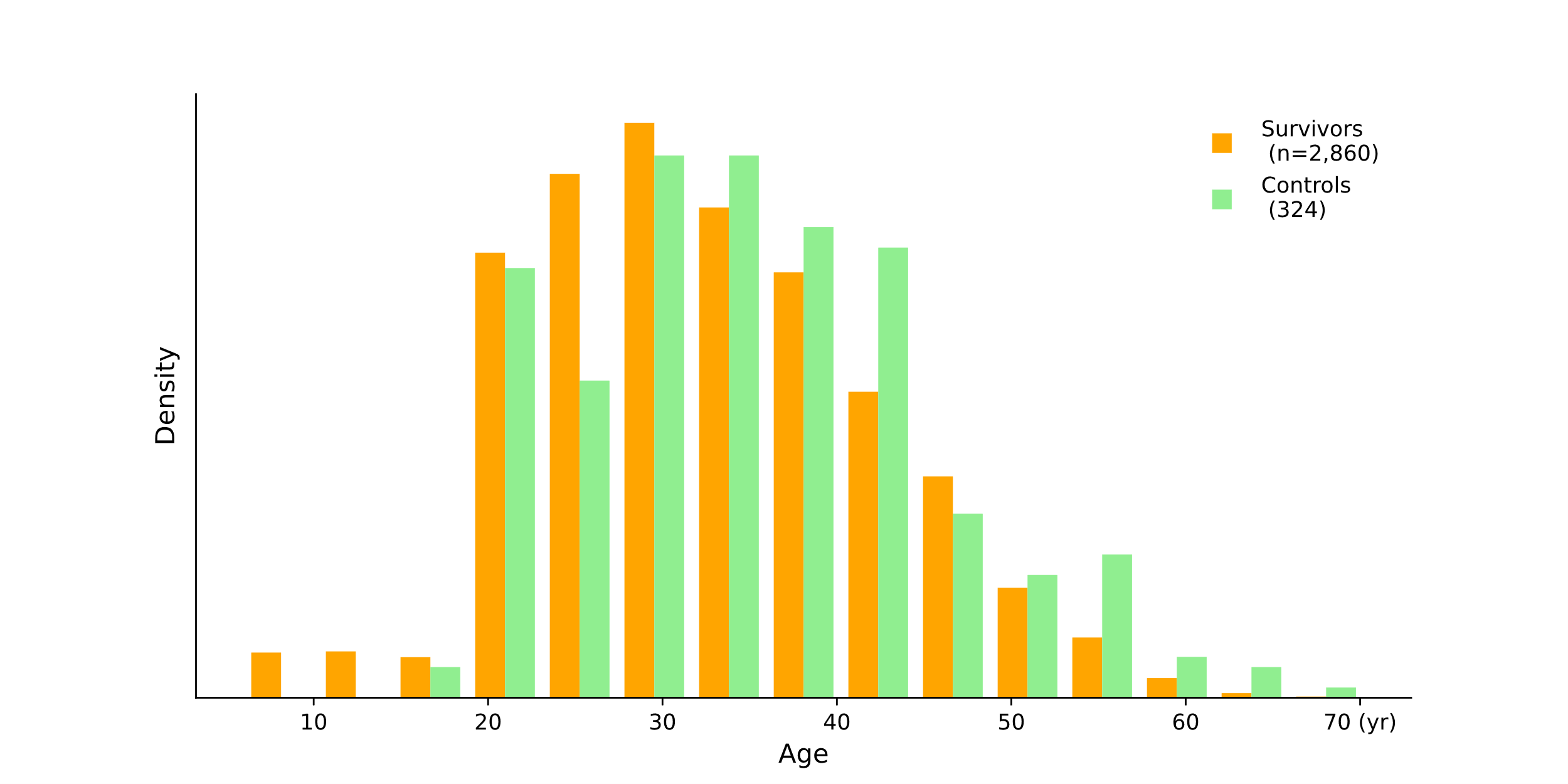
**Diagram

Description automatically generated**

The workflow documents the analysis for SNV (left flow) and indel (right). The number of variants passed each step is shown in the box.

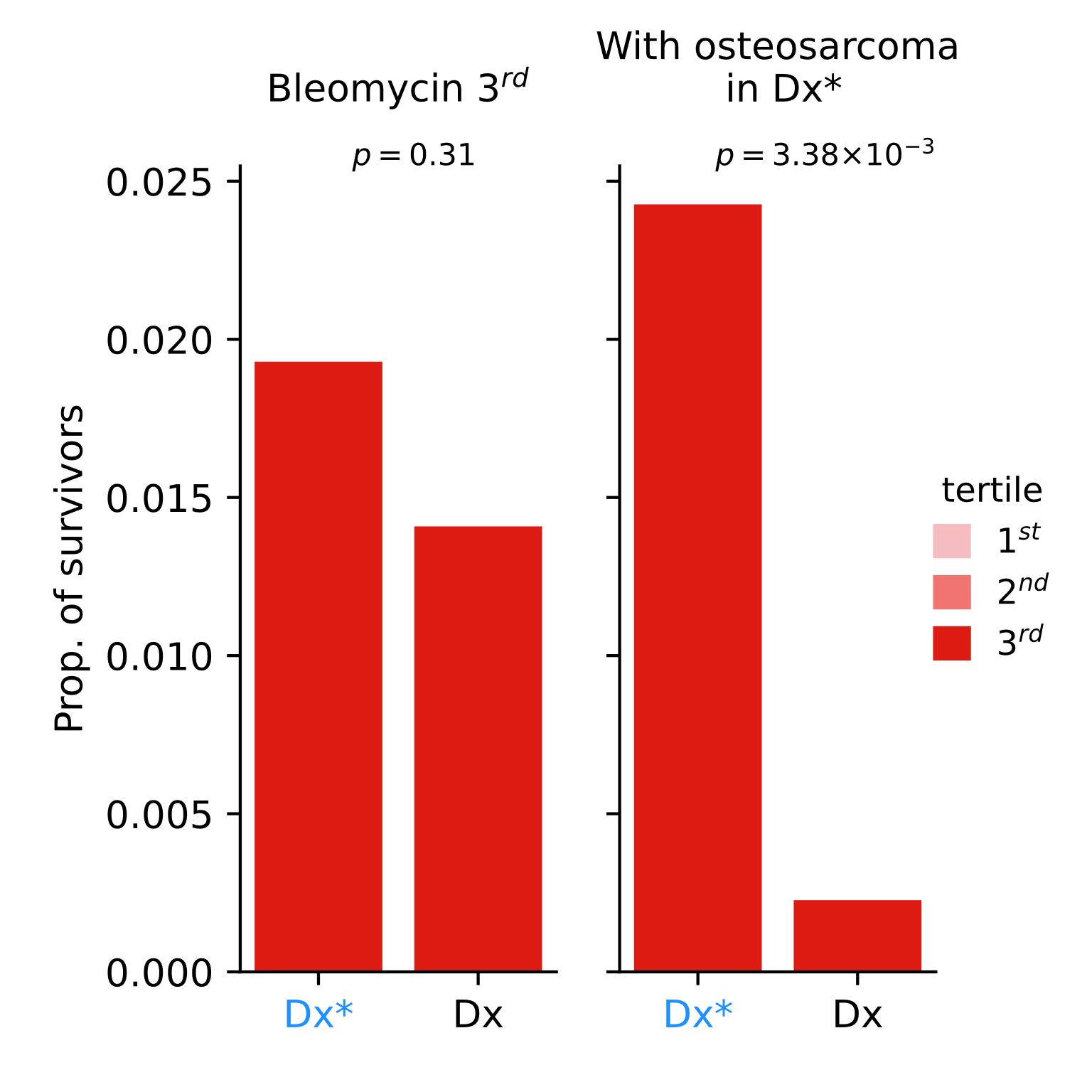
**Supplementary Figures**

**Figure S1. Age distribution of survivors and controls.**

****

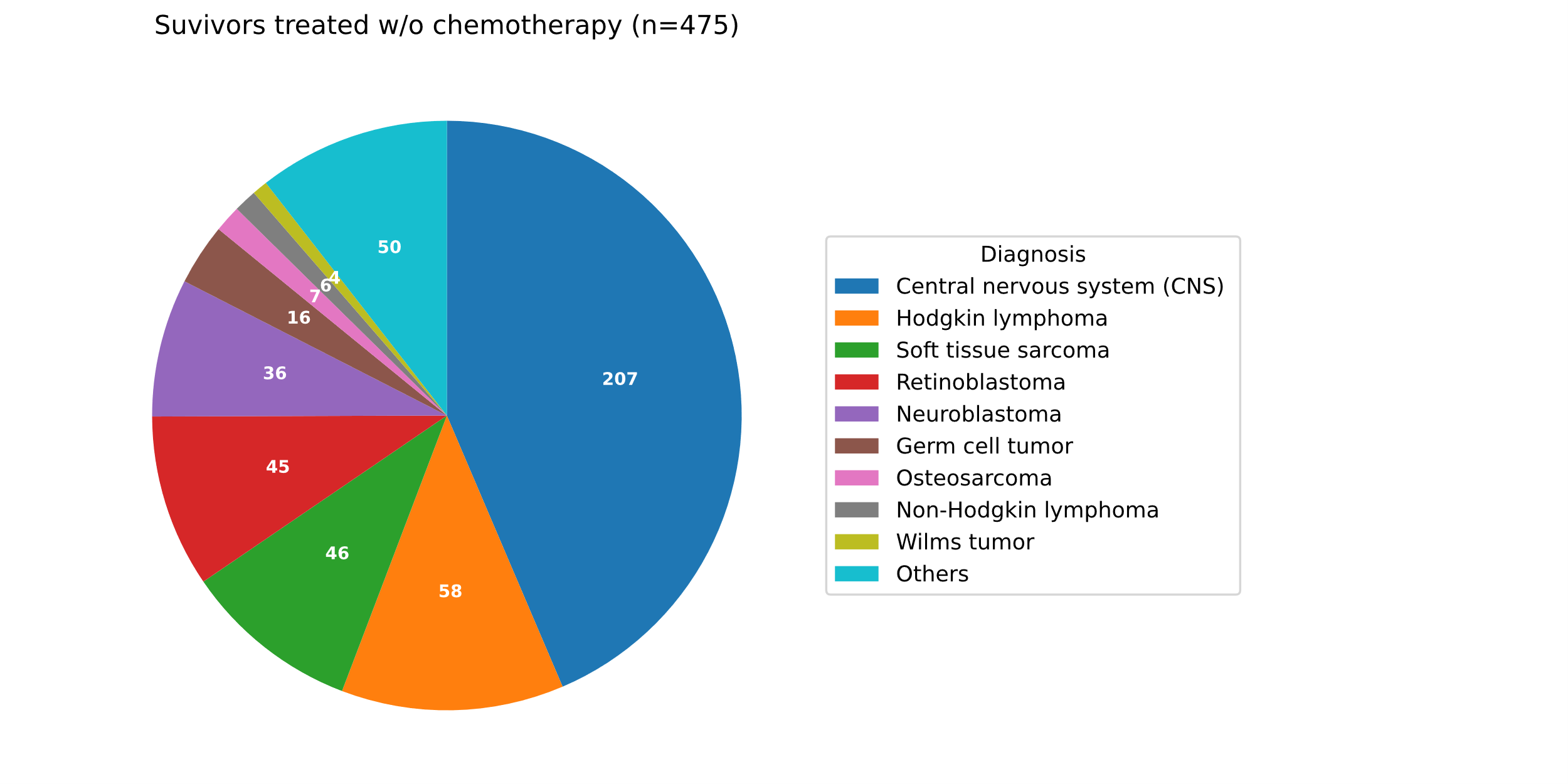
All individuals in the control cohort (green) were > 18 years old. These age distributions differed at Kolmogorov–Smirnov *p* < 1.0 × 10-16.

**Figure S2. Bleomycin usage in osteosarcoma.**

****

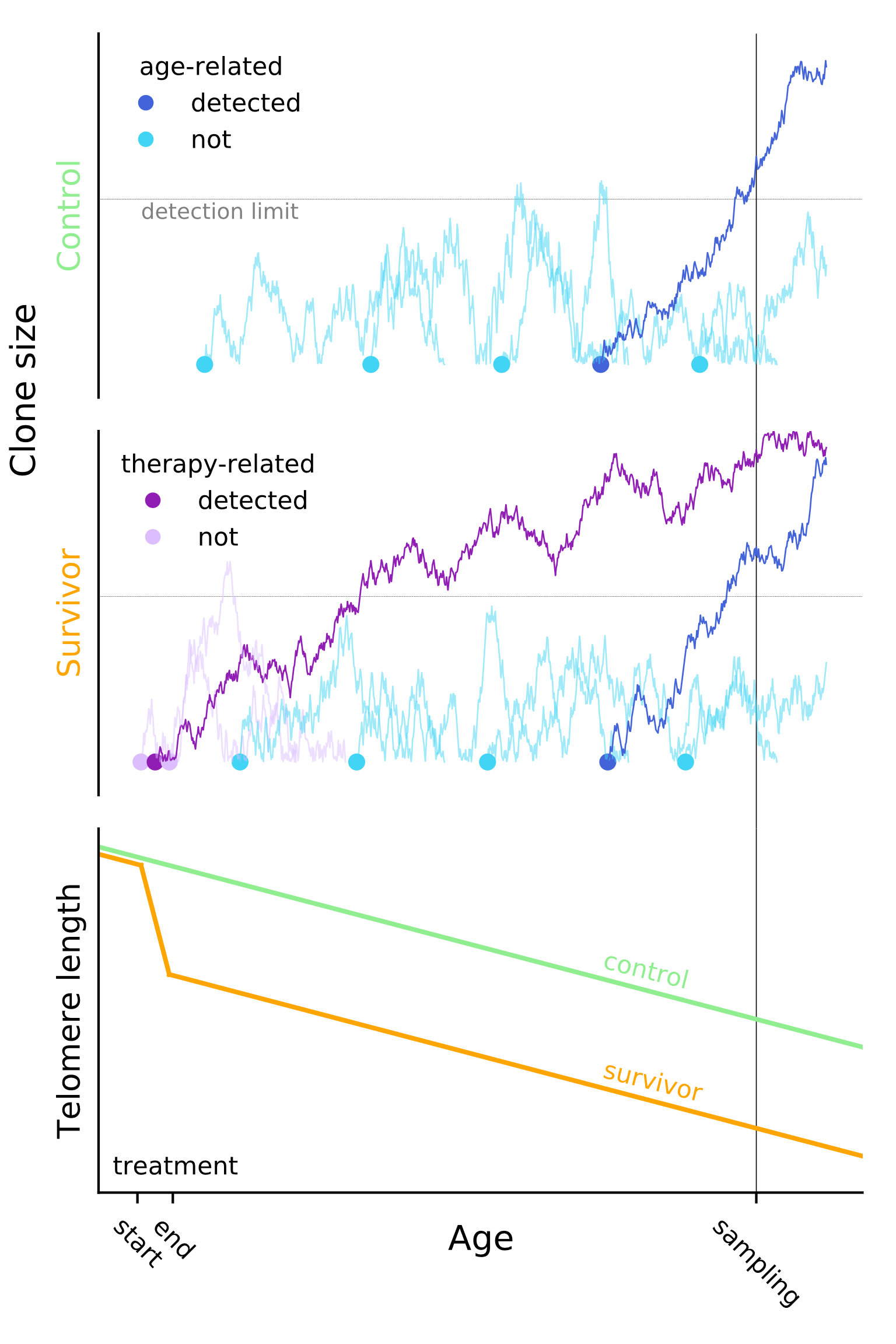
To supplement the analysis for the 2nd tertile of bleomycin (Figure 2D), the 3rd tertile was similarly plotted for *Dx\** and *Dx*. Bleomycin was most intensively used for osteosarcoma, which barely missed the *Dx\** classification (*Dx\**/*Dx* odds ratio 1.85, 95% CI: 0.98 ‒ 3.53). This characteristic bleomycin usage equated the 3rd tertile distribution between *Dx\** and *Dx* (*left*) and the ad hoc reclassification of osteosarcoma to *Dx\** created a pattern similar to the 2nd tertile (*right*).

**Figure S3. Childhood cancer types of 475 survivors who were treated without chemotherapy.**

****

Diagnoses of the survivors who had not received either of alkylating agents, platinum, anthracyclines, bleomycin, dactinomycin, epipodophyllotoxins, methotrexate, vinca alkaloids. The diagnosis class designated as *Others* (n = 50) include 12 melanoma, 8 acute lymphoblastic leukemia, 4 colon carcinoma, 3 liver malignancy, 2 chronic myeloid leukemia, 1 acute myeloid leukemia, 1 nasopharyngeal carcinoma, and 19 unspecified diseases.

**Figure S4. Conceptual model for illustrating the growth pattern of age- and therapy-related CH clones.**

****

Clone trajectories are generated as a random walk (not actual data) for illustration purpose. Age-related CH (blue) occur in HSCs and may expand over time and be detectable at the time of blood sampling (clone in darker blue). By contrast, therapy-related clones (purple) in the survivors emerged during treatment, which may also grow to a detectable size (clone in darker purple). Under the current model, age-related clones in survivors (orange) emerge at a similar frequency in controls (green) given the comparable telomere attrition rates (bottom) detected in both populations (details in Figure 1D).

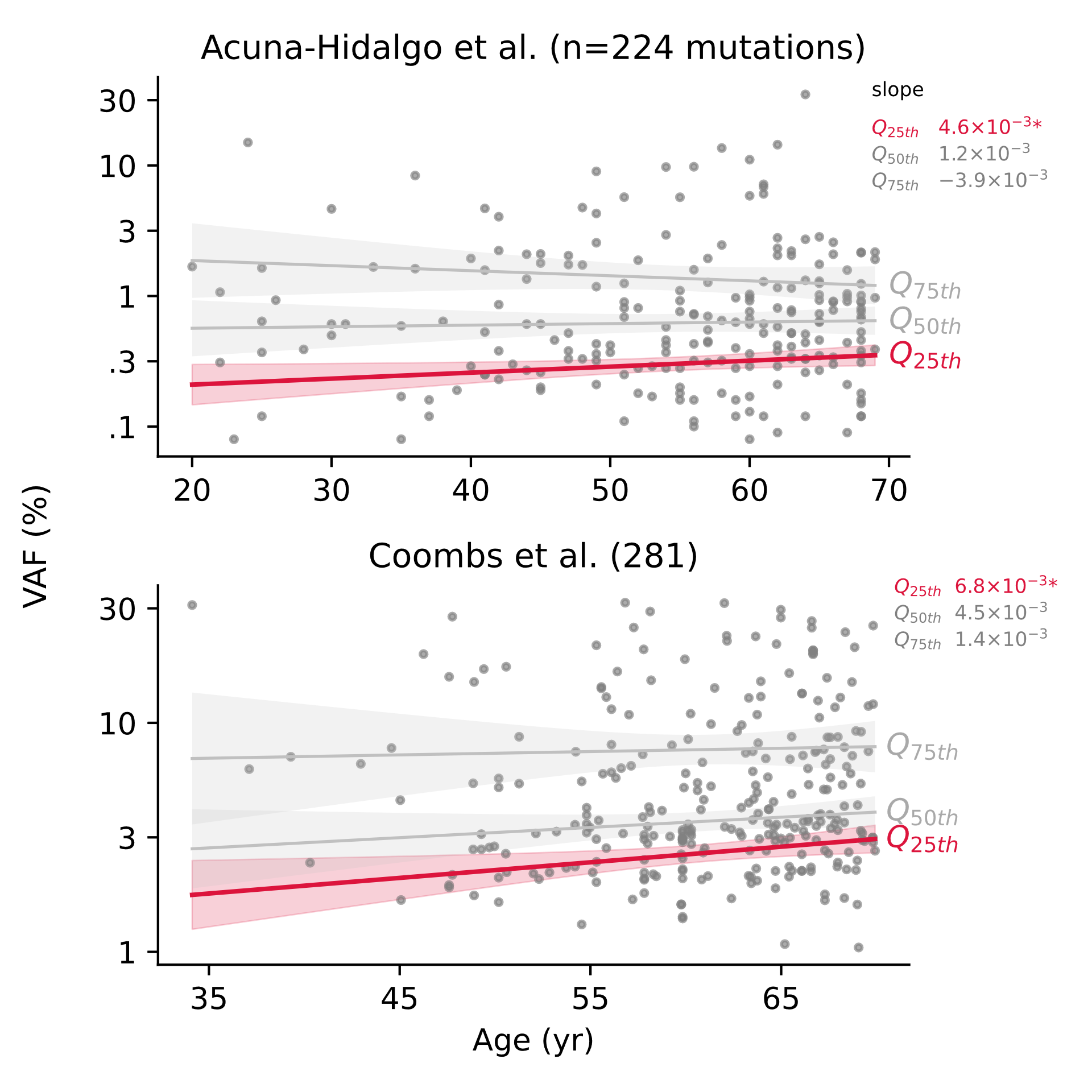
**Figure S5. Cellular distribution of *STAT3 Y640F* in blood samples from HL survivors.**

**Chart, waterfall chart

Description automatically generated**

The cell types in the peripheral blood samples were identified by Mission Bio Tapestri assay. The total number of cells profiled by this assay is labeled in parenthesis along with the sample name. The y-axis range, in log scale, is not uniform so that low cell fraction of *STAT3* Y640F in a specific cell type is visible; the lowest tick marker for CF in labeled. Across all cell types, the cellular fraction (CF) of *STAT3 Y640* mutation (first column, shown in pink) was roughly twice as the variant allele fraction (VAF) ascertained from deep sequencing (labeled in parentheses under the sample name), showing the high concordance of these two assays. The CF of individual cell types was compared to this baseline (i.e., CF of all cell types). for each cell phenotypes. Cell types with a significant enrichment of *STAT3 Y640* CF are indicated with values (proportion test p-value corrected for multiple testing). This plot uses the same data presented in Fig. 3F.

**Figure S6. Clone expansion trend analysis on public datasets.**

****

Using two independent publicly available datasets, growth trend was analyzed by quantile regression on VAF quartiles Q25th (lower 25%), Q50th (median), and Q75th(upper 25%). Acuna-Hidalgo *et al*. (10) analyzed 2,006 healthy individuals aged from 20 to 69 years old (median: 45) for CH variant with a threshold of 0.1% VAF. Coombs *et al*. (15) analyzed 5,649 cancer patients aged 0.13 to 98.7 years old (median: 58.3). In this dataset, CH mutations found in samples that had received neither of chemotherapy nor radiotherapy at the time of blood collection were used as likely age-related. Samples aged under 70 years old were included in analysis for comparison with Acuna-Hidalgo *et al*. (10) and the current study (median age: 31.6, range: 6.0 ‒ 66.4 years). \**p* < 0.05.

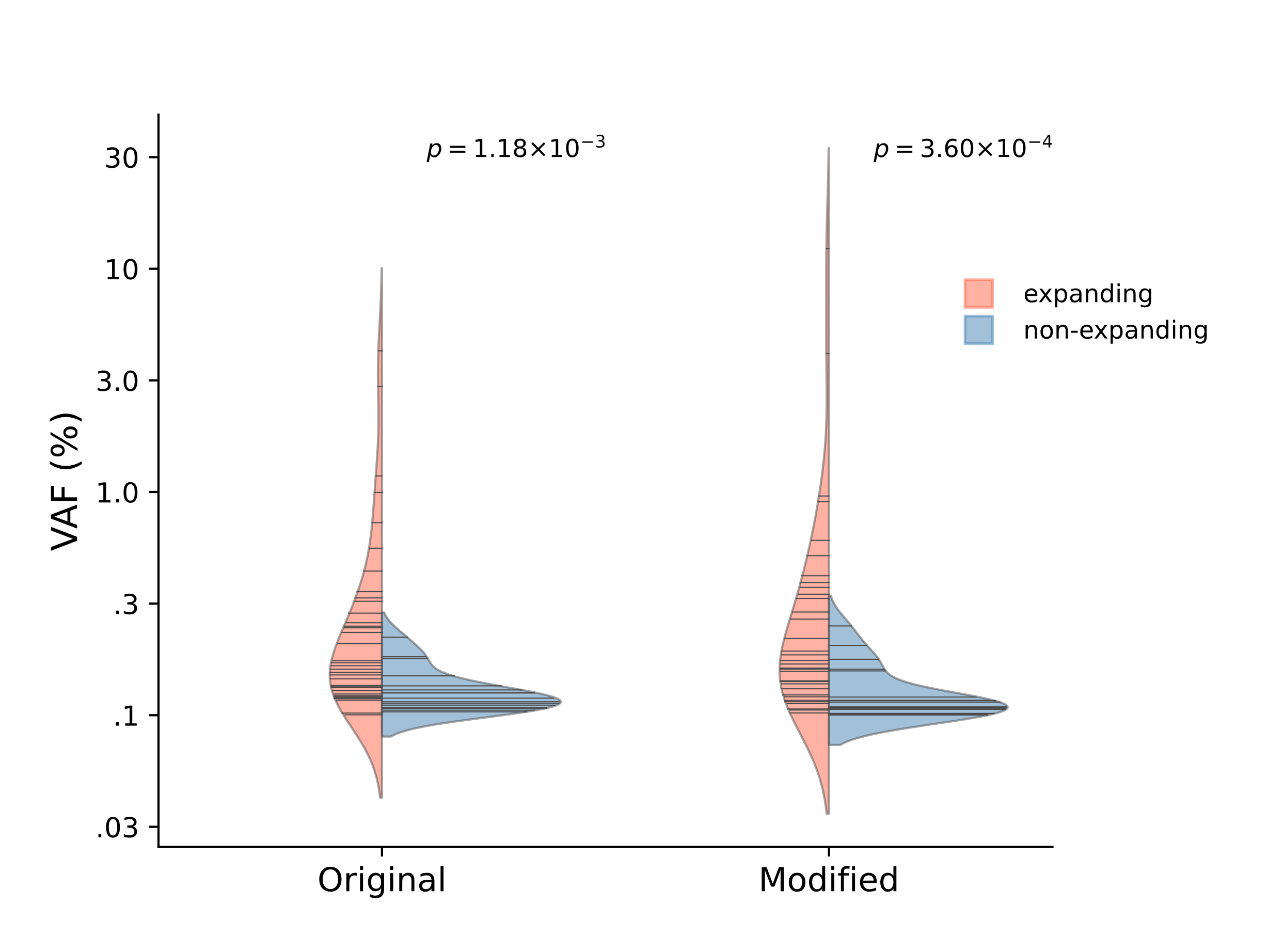
**Figure S7. Longitudinal view of serial sample analysis.**

**Chart, diagram

Description automatically generated**

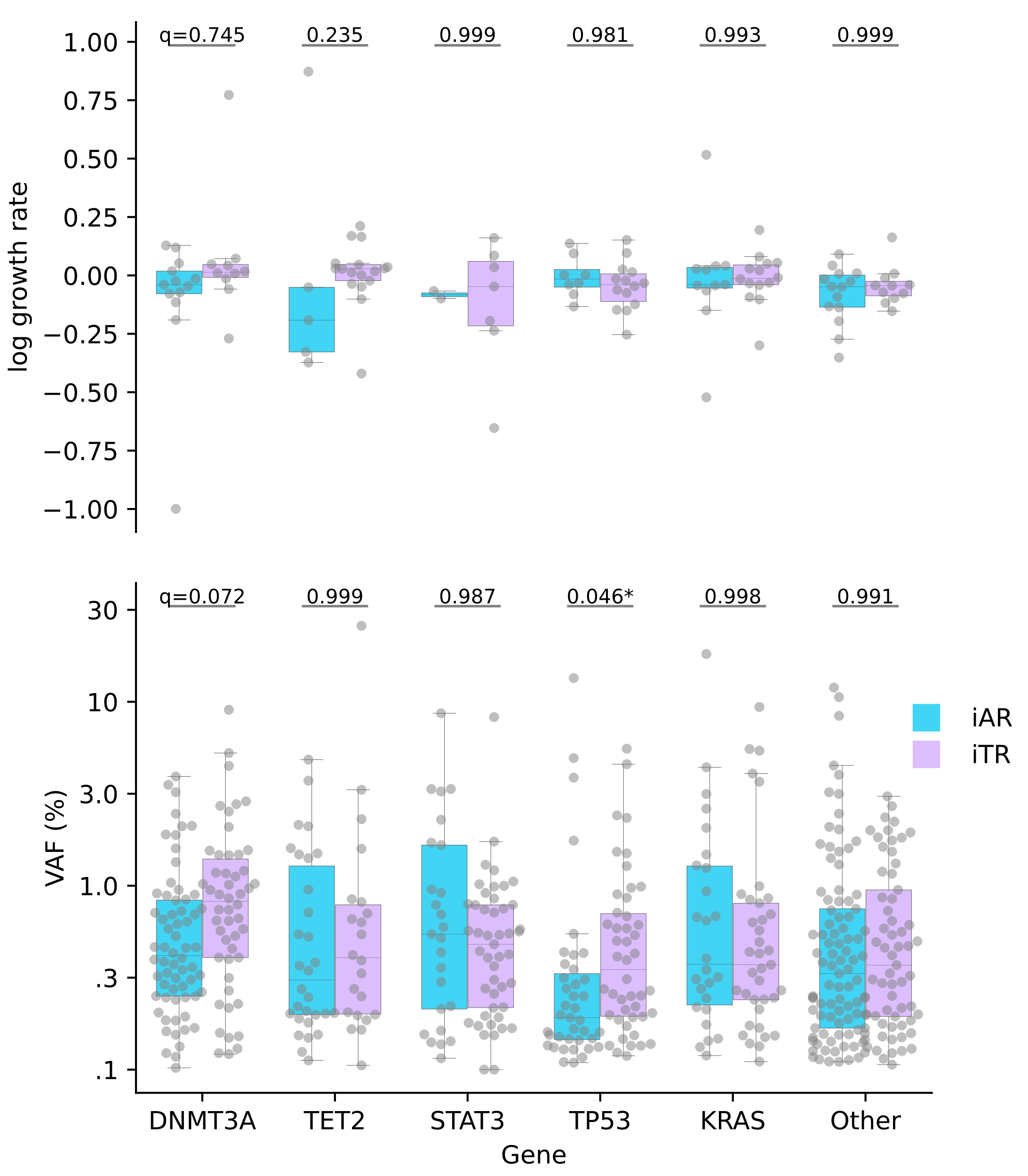
Serial observations (228 timepoints) are plotted separately for 46 iAR (*left*) and 55 iTR clones (*right*). For each clone, two consecutive timepoints are paired (connected with line): 54 pairs for iAR and 74 for iTR. For each observation pair, the VAF at the earlier timepoint was compared for the expanding (red) and non-expanding (blue) growth directions in Figure 4B (main text). The breakdown of genes (*DNMT3A*, *TET2*, *TP53*, *STAT3*, *KRAS* and Other) is also presented (bottom).

**Figure S8. Simulation of CH dynamics with clone onset manipulation.**

****

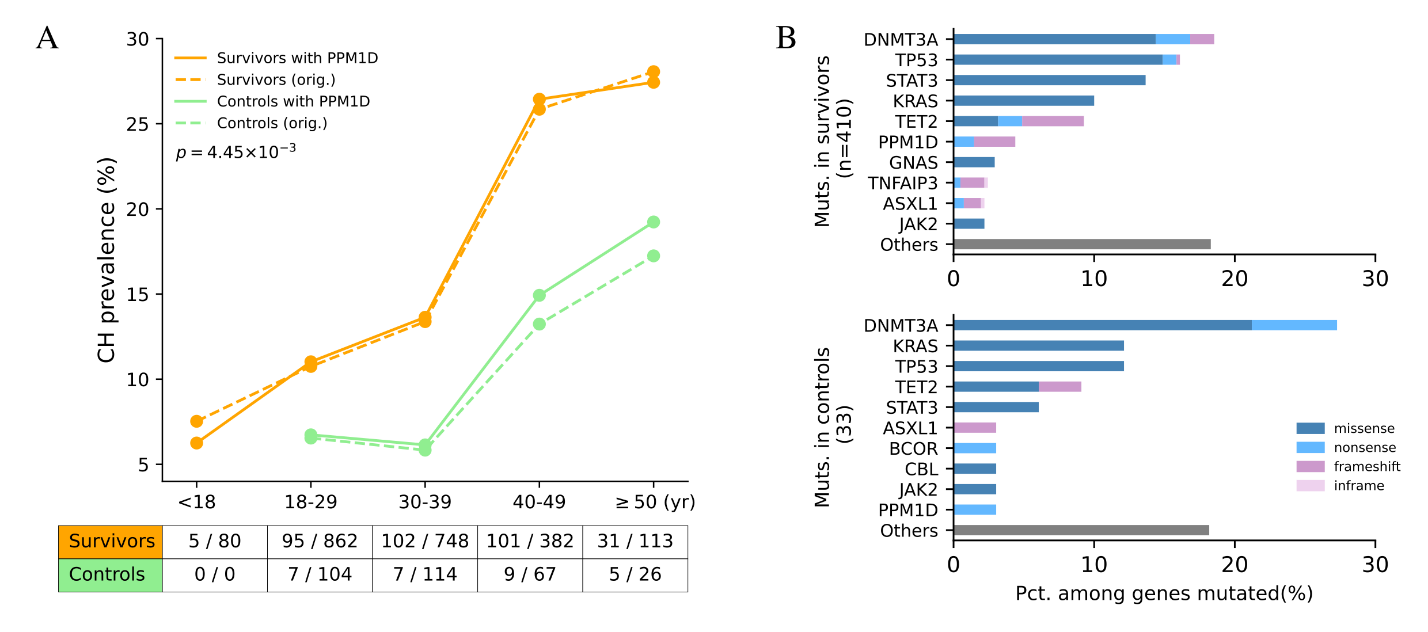
The dynamics of CH clones were simulated using the model proposed by Watson *et al.* (16). CH clones with VAF > 0.1% were sampled twice with a random interval. The clone was considered positively growing if VAF at latter time point > VAF at initial time point, otherwise as shrinking (negative growth). The original model generates CH clones at any age up to 70 years old (the max age in the current cohort). In the modified model, the mutation rate randomly elevates for a random time window that happens at 0 ‒ 20 years old to simulate a childhood cancer treatment scenario (Methods in the main text). CH trajectories started during the hyper-mutative time window was similarly sampled twice and evaluated for growth direction. Log-transformed VAFs at initial time point were compared between expanding and non-expanding growths by Mann‒Whitney test.

**Figure S9. Growth rate and size of CH clones in long-term survivors.**

****

Log-growth rate was measured as for two timepoints with using the serial samples (*top*). The growth rates between iAR and iTR were compared by Mann‒Whitney test with Šidák correction for multiple testing. Similarly, VAFs were compared using all CH mutations identified in the survivors (*bottom*). Genes less frequently mutated than *DNMT3A*, *TET2*, *STAT3*, *TP53* and *KRAS* were collectively represented as “Other”.

**Figure S10. CH prevalence re-calculated by inclusion of *PPM1D* mutations.**



In addition to the 39 genes in the original gene panel, *PPM1D* was additionally sequenced for available samples (2,185 survivors and 311 controls) which were used to recalculate CH prevalence shown in this figure. **A**, CH prevalence with *PPM1D* is plotted (*solid*): overall frequency, 15.3% for survivors and 9.3% for normal (Fisher’s exact *p* = 4.45 × 10-3). The original prevalence without *PPM1D* at each age category is also added for reference (*dashed*): overall 15.0% for survivors and 8.6% for controls. **B**, The frequency of genes mutated in survivors (*top*) and controls (*bottom*).

**Supplementary Tables**

**Table S1. 39 genes analyzed for clonal hematopoiesis by deep sequencing.**

This included genes mutated in three or more samples in Jaiswal *et al*. (7) (n=26) or cancer predisposition genes with pathogenic or likely pathogenic mutations in long-term survivors of pediatric cancer reported by Wang *et al*. (1) (n = 18). Five genes appear in both data sets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Genes selected from Jaiswal *et al.* | | | | |
| ***ASXL1*** | ***BCOR*** | *BRCC3* | *CUX1* | ***DNMT3A*** |
| ***GNAS*** | ***GNB1*** | ***IDH2*** | ***JAK2*** | *LUC7L2* |
| *KMT2D* | ***NRAS*** | *PRDM1* | ***RAD21*** | ***SF3B1*** |
| ***SRSF2*** | ***STAT3*** | ***TET2*** | *TNFAIP3* | *U2AF1* |
| *ZRSR2* |  |  |  |  |
|  |  |  |  |  |
| Genes selected from Wang *et al.* | | | | |
| *APC* | *BRCA1* | *BRCA2* | *DICER1* | *MSH6* |
| ***NF1*** | *PALB2* | *PMS2* | *PTCH1* | *RB1* |
| *SDHA* | *SUFU* | ***WT1*** |  |  |
|  |  |  |  |  |
| Genes selected from both Jaiswal *et al*. and Wang *et al.* | | | | |
| *CREBBP* | ***CBL*** | ***KRAS*** | *PTPN11* | ***TP53*** |

Bold: genes prioritized for ddPCR validation.

**Table S2. CH variants identified in survivors and controls.**

See Table S2 in “SupplementaryTablesS2\_S7\_S8\_S9.xlsx”.

**Table S3. Comparison between CH status and demographic variables.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Survivors | | | | Controls | | | |
|  | Overall (%) | CH+ | CH- | *p*-value | Overall | CH+ | CH- | *p*-value |
| #Samples | 2860 | 430 | 2430 |  | 324 | 28 | 296 |  |
| Sex | 0.994 | | | | 0.589 | | | |
| Male | 1506 (52.7) | 227 (52.8) | 1279 (52.6) |  | 152 (46.9) | 15 (53.6) | 137 (46.3) |  |
| Female | 1354 (47.3) | 203 (47.2) | 1151 (47.4) |  | 172 (53.1) | 13 (46.4) | 159 (53.7) |  |
| Race | 0.148 | | | | 0.346 | | | |
| White | 2385 (83.4) | 364 (84.7) | 2021 (83.2) |  | 292 (90.1) | 27 (96.4) | 265 (89.5) |  |
| Black | 432 (15.1) | 56 (13.0) | 376 (15.5) |  | 21 (6.5) | 0 (0.0) | 21 (7.1) |  |
| Other | 43 (1.5) | 10 (2.3) | 33 (1.3) |  | 11 (3.4) | 1 (3.6) | 10 (3.4) |  |
| Age |  |  |  | 8.2×10-19 |  |  |  | 0.023 |
| median  (range) | 31.6  (6.0 ‒ 66.4) | 36.3  (8.6 ‒ 65.9) | 31.1  (6.0 ‒ 66.4) |  | 34.6  (18.3 ‒ 70.2) | 40.2  (19.9 ‒ 70.2) | 34.3  (18.3 ‒ 65.9) |  |

*p*-valuesare based on χ2 test between CH-positive (CH+) and CH-negative (CH) individuals, except for age, which is a Mann‒Whitney comparison.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | | | |
|  | Overall (%) | Carrier | Non-carrier | *p*-value |
| CPG a | 0.078 | | | |
| CH+ | 430 (15.0) | 10 (8.9) | 420 (15.3) |  |
| CH- | 2430 (85.0) | 102 (91.1) | 2328 (84.7) |  |
| DDR b | 0.417 | | | |
| CH+ | 430 (15.0) | 20 (17.9) | 410 (14.9) |  |
| CH- | 2430 (85.0) | 92 (82.9) | 2338 (85.1) |  |
| RT-naïve DDR c |  |  |  | 0.0496 |
| CH+ | 164 (12.1) | 11 (21.2) | 153 (11.7) |  |
| CH- | 1195 (87.9) | 41 (78.8) | 1154 (88.3) |  |
| RT-exposed DDR d |  |  |  | 0.730 |
| CH+ | 266 (17.7) | 9 (15.0) | 257 (17.8) |  |
| CH- | 1235 (82.3) | 51 (85.0) | 1184 (82.2) |  |
| Chemo-naïve DDR e |  |  |  | 0.634 |
| CH+ | 41 (9.1) | 2 (13.3) | 39 (8.8) |  |
| CH- | 419 (90.9) | 13 (86.7) | 406 (91.2) |  |
| Chemo-exposed DDR f |  |  |  | 0.485 |
| CH+ | 389 (16.2) | 18 (18.6) | 371 (16.1) |  |
| CH- | 2011 (83.8) | 79 (81.4) | 1932 (83.9) |  |

**Table S4. Association between CH and pathogenic germline mutations in survivors**

a Germline mutation of 60 autosomal dominant cancer predisposition genes (CPG) as defined in Wang *et al.* (1). Mutations classified as pathogenic were used for analysis. b Germline mutation of 127 DNA damage repair (DDR) genes curated in Qin *et al.* (17). Mutations classified as pathogenic were used for analysis. c, d The association between CH and DDR mutation status was separately analyzed in non-irradicated (naïve) and irradicated (exposed) individuals. e, f Survivors were similarly analyzed after stratified by exposure to any of the following chemotherapy (Chemo) agents: alkylating agents, platinum, anthracyclines, bleomycin, dactinomycin, epipodophyllotoxins, methotrexate, vinca alkaloids. *p*-valuesrepresent Fisher’s exact comparison.

**Table S5.** **The distribution *STAT3* mutations in Hodgkin Lymphoma (HL) survivors and other survivors**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | #HL survivors | | #Other survivors | | *p*-value b |
| w/ mut. | | 33 a | | 34 | 9.21 × 10-14 |
| w/o | | 323 | | 2,470 |
|  | | | | | | |
| *STAT3* AA Change c | | #HL survivors | | #Other survivors | *p*-value d |
| S614R (TGC>TCC) | | 0 | | 3 | 1.0 |
| G618R (CCT>CGT) | | 4 | | 9 | 0.081 |
| Y640F (GTA>GAA) | | 24 | | 12 | 5.62×10-11 |
| N647I (GTT>GAT) | | 4 | | 3 | 0.012 |
| D661V (ATC>AAC) | | 1 | | 3 | 0.392 |
| D661Y (TCC>TAC) | | 3 | | 4 | 0.058 |
|  | | | | | | |
| G/C[T>A]N e | | #mut. in HL survivors | | #mut. in other survivors | *p*-value f |
| True | | 28 | | 15 | 0.006 |
| False | | 8 | | 19 |

a Three HL survivors (SJHL041530, SJHL041541, and SJHL042024) had two *STAT3* mutations each. b, f Fisher’s exact comparison between HL and other survivors. c *STAT3* mutation denoted by amino acid (AA) change with trinucleotide context change in parentheses. d Binomial comparison for the mutation frequency in HL survivors and other survivors in our cohort. e True if the trinucleotide mutation context matches the predominant pattern of COSMIC SBS25 signature which involves thymine (T) to adenine (A) change with a preceding guanine (G) or cytosine (C) and a variable trailing base (N).

**Table S6. The distribution mutational types and procarbazine usage in HL survivors.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Overall (%) | Exposed a | Not | *p*-value |
| CH pos with G/C[T>A]N b |  |  |  | 5.55×10-6 |
| True | 28 (7.9) | 26 (13.5) | 2 (1.2) |  |
| False | 328 (92.1) | 166 (86.5) | 162 (98.8) |  |
|  |  |  |  |  |
| CH pos w/o G/C[T>A]N c |  |  |  | 0.893 |
| True | 69 (19.4) | 38 (19.7) | 31 (18.9) |  |
| False | 287 (80.6) | 154 (80.3) | 133 (81.1) |  |

a, Prior exposure to procarbazine. b, True if the survivor is CH positive with mutation(s) of T>A with preceding G/C. False if CH-negative or CH-positive but with different pattern. c, True if the survivor is CH positive with mutations (s) other than T>A with preceding G/C. False if CH-negative or CH-positive with mutation(s) of T>A with preceding G/C. *p*-values represent Fisher’s exact comparison.

**Table S7. CH variants in serial analysis.**

See Table S7 in “SupplementaryTablesS2\_S7\_S8\_S9.xlsx”.

**Table S8. Demographic and clinical variables.**

See Table S8 in “SupplementaryTablesS2\_S7\_S8\_S9.xlsx”.

**Table S9. Primer and probes used for ddPCR.**

See Table S9 in “SupplementaryTablesS2\_S7\_S8\_S9.xlsx”.

**Supplementary References**

1. 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.

2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**;25:1754-1760.

3. 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.

4. Liu FT, Ting KM, Zhou ZH. Isolation forest. In 2008 eighth IEEE international conference on data mining **2008** Dec 15 (pp. 413-422). IEEE.

5. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **2010**;38:e164. doi: 10.1093/nar/gkq603.

6. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, *et al*. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **2019**;47(D1):D941-D947.

7. 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.

8. 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.

9. Zink F, Stacey SN, Norddahl GL, Frigge ML, Magnusson OT, Jonsdottir I, *et al*. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood*. **2017**;130:742-752. doi: 10.1182/blood-2017-02-769869.

10. 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.

11. 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.

12. Hagiwara K, Edmonson MN, Wheeler DA, Zhang J. indelPost: harmonizing ambiguities in simple and complex indel alignments. *Bioinformatics* **2022**;38:549-551.

13. 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.

14. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* **2002**;12:656-664.

15. 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.

16. 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.

17. 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.