SUPPLEMENTARY INFORMATION: GENOMIC DETERMINANTS OF OUTCOME IN ACUTE LYMPHOBLASTIC LEUKEMIA

Ti-Cheng Chang Ph.D.[1*], Wenan Chen, Ph.D[1*†], Chunxu Qu, Ph.D.[2], Zhongshan Cheng Ph.D.[1], Dale Hedges Ph.D.[1], Abdelrahman Elsayed Ph.D.[3], Stanley B. Pounds Ph.D[3], Mary Shago Ph.D.[4], Karen R. Rabin M.D., Ph.D[5], Elizabeth A. Raetz M.D.[6], Meenakshi Devidas, Ph.D.[7], Cheng Cheng Ph.D[3], Anne Angiolillo M.D.[8], Pradyuamma Baviskar Ph.D.[2], Michael Borowitz M.D.[9], Michael J. Burke M.D.[10], Andrew Carroll Ph.D.[11], William L. Carroll M.D.[6], I-Ming Chen D.V.M[12], Richard Harvey Ph.D[12], Nyla Heerema Ph.D[13], Ilaria Iacobucci, Ph.D.[2], Jeremy R. Wang Ph.D[14], Sima Jeha, M.D.[15], Eric Larsen M.D.[16], Leonard Mattano M.D. [17], Kelly Maloney M.D.[18], Ching-Hon Pui, M.D.[15], Nilsa C, Ramirez[19], Wanda Salzer M.D.[20], Cheryl Willman M.D.[21] Naomi Winick M.D.[22], Brent Wood M.D.[23], Stephen P. Hunger M.D.[24], Gang Wu Ph.D.[1,2], Charles G. Mullighan, M.B.,B.S., M.D.[2**], Mignon L. Loh M.D.[25**]

[1]Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis TN, USA
[2]Department of Pathology, St. Jude Children's Research Hospital, Memphis TN, USA
[3]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis TN, USA
[4]Department of Laboratory Medicine and Pathobiology, University of Toronto, CA
[5]Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA
[6]Department of Pediatrics, Perlmutter Cancer Center, NYU Langone Hospital, NY, NY, USA
[7]Global Pediatric Medicine, St. Jude Children's Research Hospital, Memphis TN, USA
[8]Children's National Medical Center, Washington DC, USA
[9]Department of Pathology, Johns Hopkins University, Baltimore MD, USA
[10]Division of Pediatric Hematology-Oncology, Medical College of Wisconsin, Milwaukee WI, USA
[11]Department of Genetics, University of Alabama at Birmingham, Birmingham AL, USA
[12]Department of Pathology, University of New Mexico, Albuquerque, NM, USA
[13]The Ohio State University, Columbus OH, USA
[14]Department of Genetics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill NC, USA
[15]Department of Oncology, St. Jude Children's Research Hospital, Memphis TN, USA
[16]Department of Pediatrics, Maine Children's Cancer Program, Scarborough ME, USA
[17]HARP Pharma Consulting, Mystic CT, USA
[18]Department of Pediatrics and Children's Hospital Colorado, University of Colorado, Aurora CO, USA
[19]Institute for Genomic Medicine and Biopathology Center, Nationwide Children's Hospital, Departments of Pathology and Pediatrics, Ohio State University, Columbus, OH 43205, USA
[20]Uniformed Services University, School of Medicine, Bethesda, MD, USA
[21]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA
[22]Department of Pediatric Hematology Oncology and Simmons Cancer Center, University of Texas Southwestern Medical Center, Dallas TX, USA
[23]Department of Pathology and Laboratory Medicine, Children's Hospital Los Angeles, University of Southern California, Los Angeles, CA, USA
[24]Department of Pediatrics and the Center for Childhood Cancer Research, Children's Hospital of Philadelphia, and the Perelman School of Medicine at the University of Pennsylvania, Philadelphia PA, USA
[25]Department of Pediatrics and the Ben Towne Center for Childhood Cancer Research, Seattle Children's Hospital, University of Washington, Seattle, WA, USA

[†]Current affiliation: Division of Computational Biology, Mayo Clinic, Rochester, MN, USA
*Contributed equally; **Contributed equally to senior authorship

TABLE OF CONTENTS

SUPPLEMENTARY METHODS

Patients/specimen characteristics

Patients were considered for this study if they were eligible, evaluable, and enrolled on one of four clinical trials with mature outcomes: AALL0331 (NCT00103285)[1], AALL0932 (NCT01190930)[2,3], AALL0232 (NCT00075725)[4,5], and AALL1131 (NCT02883049)[6]. AALL0331 enrolled NCI SR patients and randomized favorable risk patients to intensive asparaginase vs not, average risk patients to intensive consolidation versus standard consolidation, and nonrandomly assigned high risk patients (slow early responders) to full augmented BFM based therapy. Patients were considered to be rapid early responders (RER) if they achieved a day 8 M1 marrow and Day 29 MRD < .1% by flow cytometry. AALL0932 enrolled NCI SR patients and randomized favorable risk patients to P9904 versus standard of care CCG based therapy. Average risk patients were randomized in a 2 x 2 fashion during maintenance therapy to higher dose oral methotrexate vs standard oral dosing, and q 12 weeks pulses of vincristine/dexamethasone versus q 4 week pulses. Rapid early responders needed to achieve an EOI MRD < 0.01% by flow cytometry. AALL0232 enrolled NCI HR patients and randomized RER patients in Interim maintenance 1 between high dose methotrexate and Capizzi methotrexate. There was an additional steroid randomization during induction. Patients were considered to be RER if they achieved a day 8 M1 marrow and Day 29 MRD < .1% by flow cytometry. Slow early responders received full augmented BFM therapy with a double delayed intensification. AALL1131 randomized HR patients between single versus triple intrathecal therapies. Rapid early responders needed to achieve an EOI MRD < 0.01% by flow cytometry. Of note, patients initially enrolled on AALL0932 who did not achieve a rapid early response were eligible to roll over to the high risk randomized question on AALL1131. NCI HR patients who were slow early responders were randomized to more intensive consolidative strategies that initially included clofarabine but subsequently did not, due to the excessive toxicity of clofarabine.

Between January 1, 2004 and December 31, 2013, a total of 7290 NCI SR patients with either favorable or neutral cytogenetics were enrolled (Figure S2). Of 1653 selected patients, 495 experienced relapse and 1158 did not. After assessing for adequate sample collection at diagnosis and at end induction (germline), and performing the assays described below, a total of 1381 patients were analyzed (405 with relapse, 976 without). For NCI HR patients, the analyzed sample group was 115 (34 relapsed, 81 was not). Data and remission status was current to 3/31/21. The MRD status of these NCI SR and HR patients is shown in the CONSORT diagrams. All patients and/or their guardians consented to the collection of research specimens according to the Declaration of Helsinki. The clinical trials were approved by the central IRB of the National Institutes of Health. Among the total 1496 analyzed samples, 1,462 had both WGS and WTS, 31 had WGS only, and 3 had only WTS. Of the 1,493 WGS, 1,472 also had WES. We observed differences in clinical and genetic features between patients studied compared to those not studied attributable to the intentional enrichment of patients that experienced relapse; these included age (P=0.013), *ETV6::RUNX1* (P=0.0001), high hyperdiploidy (P<0.0001), CNS status (P=0.046) and MRD status at day 8 (P=0.041; Table S3). To validate associations of chromosomal alterations and outcome in hyperdiploid ALL, we studied 267 patients from St. Jude Total Therapy 15 and 16 cohorts.[7,8]

Study design

We focused on standard-risk patients from two trials, AALL0331 and AALL0932, and included a small number of high-risk B-ALL patients with favorable cytogenetic features included from two high risk trials AALL0232 and AALL1131. Because NCI SR ALL patients generally have lower relapse rates than others, we wanted to increase the possibilities of identifying relapse predictors within the MP2PRT study size limit of 1500 individuals. Therefore, we used a case-control design in which cases were defined as all those who relapsed, and controls included patients who remained in continuous complete remission for at least 5 years at the time a snapshot of study data was frozen (March 31, 2021). The cases to control ratio was set at about 1:2. and the

distribution of MRD status at the end of induction (<0.01% vs. ≥ 0.01%) was set to be similar between selected cases and controls within each study (Supplementary Figure S2). Because MRD status was matched in our design, we only included it as a covariate for the association analysis between clinical features, subtypes, and relapse risk. We did not include MRD status as a covariate in the genome-wide association analysis between genetic lesions and relapse risk.

Choice of analysis methods

A previous study reported good performance of logistic regression for similar "extreme" case control designs for time-to-event data[9], which we confirmed by performing a simulation study. We thus used logistic regression to scan the observed genomic lesions for their associations with relapse, with the awareness that the odds ratios could be biased due to the study design. We believe this is reasonable for this part of the analysis because our goal herein was to detect associations instead of quantifying the degree of association. For analysis illustrated by Kaplan-Meier plots, we incorporated the sampling probability to better reflect the time-to-event outcome.

Biospecimen pathology qualification and nucleic acid extraction

The COG selected 1,801 biospecimens that met the study criteria, of which 128 cases were housed at the cell bank at the University of New Mexico (UNM) and the remainder were from the Biopathology Center (BPC) at Nationwide Children's Hospital (BPC). The BPC processed leukemia samples with germline controls from a total of 1,602 cases, of which 1,513 cases qualified for sufficient material and were sent for genomic analysis. Cases were disqualified due to insufficient blasts (<70%), RNA integrity scores of <7.0, insufficient nucleic acid yield, and/or genotypic mismatch between the tumor and germline samples. There were 105 cases with low blast content, of which 63 cases were salvaged by performing Florescence Activated Cell Sorting (FACS) before the nucleic acid extraction.

Of note, hematopathology quality control using Wright/Giemsa stained material was performed on each leukemia specimen from either a cytospin or smear prepared by the BPC upon

receipt of the specimens. The percent blast and other pathology annotations were also assessed. Leukemia samples with ≥70% blasts were submitted for nucleic acid extraction.

DNA and RNA were extracted, and quality was assessed at the BPC. RNA and DNA were extracted from tumor using a modification of the DNA/RNA AllPrep kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a *mir*Vana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp DNA Blood Midi kit (Qiagen). RNA and DNA were already extracted for 3.9% of the tumor cases. For these samples, the RNA samples were extracted using the Invitrogen Life Technologies TRIzol reagent and the DNA using QiaAmp DNA mini kit (Qiagen). A custom Sequenom SNP panel was utilized to verify that tumor DNA and germline DNA representing a case were derived from the same patient. RNA was analyzed via the RNA6000 Nano assay (Agilent) for determination of an RNA Integrity Number (RIN). Cases yielding 1.5 µg tumor DNA (0.75 µg minimum), 1.5 µg tumor RNA, 1.5 µg of germline DNA (0.75 µg minimum), and 0.5 µg germline RNA were included in this study.

To enrich samples with low blast enumeration, fluorescence activated cell sorting was performed on a BD Influx sorter using BD Sortware software. The sorter was run with a sheath pressure of 27 using a 100 uM nozzle. Cells were collected in chilled 12 x 75 FACS tubes using a four-way sort. All post-sort samples were run for a purity check. Gating strategy involved initial gate of viable cells based on scatter followed by a singlets gate and then a 7AAD Live/Dead stain. Blasts and lymphocytes were separated based on CD45 labeling with further separation of lymphocytes into T-cell and B-cell populations based on CD7 and CD19 positivity respectfully. Blasts were further characterized from the CD45 negative population with CD10.

Exome sequencing

Library Construction.

An aliquot of genomic DNA (100-150ng in 50µL) was used as the input into DNA fragmentation (aka shearing). Shearing was performed acoustically using a Covaris focused ultrasonicator,

targeting 150bp fragments. Library preparation was performed using a commercially available kit provided by KAPA Biosystems (KAPA HyperPrep Kit with Library Amplification product KK8504) and IDT's duplex UMI adapters. Unique 8-base dual index sequences embedded within the p5 and p7 primers (purchased from IDT) were added during PCR. Enzymatic clean-ups were performed using Beckman Coulter AMPure XP beads with elution volumes reduced to 30μL to maximize library concentration.

Following library construction, library quantification was performed using the Invitrogen Quant-It broad range dsDNA quantification assay kit (Thermo Scientific Catalog: Q33130) with a 1:200 PicoGreen dilution. Following quantification, each library was normalized to a concentration of 35 ng/μL, using Tris-HCl, 10mM, pH 8.0. All steps performed during the library construction process and library quantification process were performed on the Agilent Bravo liquid handling system.

In-solution hybrid selection.

After library construction, hybridization and capture were performed using the relevant components of IDT's XGen hybridization and wash kit and following the manufacturer's suggested protocol, with several exceptions. A set of 12-plex pre-hybridization pools were created. These pre-hybridization pools were created by equivolume pooling of the normalized libraries, Human Cot-1 and IDT XGen blocking oligos. The pre-hybridization pools undergo lyophilization using the Biotage SPE-DRY. Post lyophilization, custom exome bait (TWIST Biosciences) along with hybridization mastermix was added to the lyophilized pool prior to resuspension and incubated overnight. Library normalization and hybridization setup are performed on a Hamilton Starlet liquid handling platform, while target capture is performed on the Agilent Bravo automated platform. Post capture, a PCR was performed to amplify the capture material.

Preparation of libraries for cluster amplification and sequencing.

After post-capture enrichment, library pools were quantified using qPCR (automated assay on the Agilent Bravo), using a kit purchased from KAPA Biosystems with probes specific to the ends of

the adapters. Based on qPCR quantification, pools were normalized using a Hamilton Starlet to the required loading concentration. Up to 24 samples were sequenced per lane on Illumina's NovaSeq S4 sequencing technology.

Cluster amplification and sequencing.

Cluster amplification of library pools was performed according to the manufacturer's protocol (Illumina) using Exclusion Amplification cluster chemistry and NovaSeq S4 flowcells. Flowcells were sequenced on Sequencing-by-Synthesis chemistry for NovaSeq S4 flowcells using paired 151bp runs.

Human Whole Genome Sequencing

Preparation of libraries for cluster amplification and sequencing.

PCR Plus (v1.1-v1.3) was used for 29 samples due to low input DNA availability, and PCR Free methods for 1475 cases. Following fragmentation, which was identical to the exome preparation above, additional size selection was performed using a SPRI cleanup. Library preparation was performed using a commercially available kit provided by KAPA Biosystems (PCR Plus: KAPA Hyper Prep with Library Amplification Primer Mix, product KK8504; PCR Free: KAPA Hyper Prep without amplification module, product KK8505), and with palindromic forked adapters using unique 8-base index sequences embedded within the adapter (purchased from Roche). For the samples with low input DNA available, the libraries were then amplified by 10 cycles of PCR. Following sample preparation, libraries were quantified using quantitative PCR (kit purchased from KAPA Biosystems) with probes specific to the ends of the adapters. Based on qPCR quantification, libraries are normalized to 2.2nM and pooled into 24-plexes.

Cluster amplification and sequencing (NovaSeq 6000).

Sample pools were combined with NovaSeq Cluster Amp Reagents DPX1, DPX2 and DPX3 and loaded into single lanes of a NovaSeq 6000 S4 flowcell cell using the Hamilton Starlet Liquid Handling system. Cluster amplification and sequencing occurred on NovaSeq 6000 Instruments utilizing sequencing-by-synthesis kits to produce 151bp paired-end reads.

Transcriptome sequencing

RNA sequencing was performed in 1465/1496 (97.9%) cases with 1372 yielding adequate transcriptome data. Illumina Stranded Total RNA Prep with RiboZero Gold barcoded with individual tags was used for RNAseq library construction following the manufacturer's instructions (Illumina, Inc. San Diego, CA). Libraries were prepared and then pooled on an automated liquid handling systems to minimize variance. Typically, these were pools of 38-92 samples, depending on available capacity on a sequencer. Quality control was performed at every step. Libraries were quantified for concentration, fragment size and distribution using a TapeStation system. As needed, pool balance and library quality were assessed using a miSeq Nano single-end 50bp sequencing.

Total RNA-seq Library Construction for Low Input Samples.

A subset of samples did not have sufficient RNA (<250 ng) for standard RNAseq library preparation. For these, a similar approach that requires less input material was used: Illumina Stranded Total RNA Prep with RiboZero Plus. These two approaches are highly similar but differ in their depletion strategy for rRNA removal. RiboZero Plus used an enzymatic depletion strategy, whereas RiboZero Gold uses bead linked capture approach. While this difference does introduce some bias into the RNA profiles, the initial quality of the analyte appears to be the major determinant of the quality of the resulting transcriptome data. As above, after quantification, libraries were prepared according to manufacturer's instructions. Pooling, library QA/QC, and sequencing were as described above.

Total RNA Sequencing.

Indexed libraries were prepared and run on an Illumina NovaSeq 6000, paired end 100 base pairs to generate a minimum of 150 million reads per sample library with a target of greater than 90% mapped reads. In all but a few cases, all data was from the same sequencing run. A few samples needed additional read depth, which was provided with a secondary sequencing run. The raw Illumina sequence data were demultiplexed and converted to fastq files; adapter and low-quality

sequences were quantified. Samples were assessed for informational quality by mapping reads to human genome reference (hg38), estimating total number of reads that mapped, amount of RNA mapping to coding regions, amount of rRNA in sample, number of genes expressed, and relative expression of housekeeping genes. The 1480 samples passing this QAQC were then clustered with other expression data from similar and distinct tumor types to confirm expected expression patterns. Atypical samples were SNP typed to confirm source analyte. FASTQ files of all reads were then uploaded to the GDC repository and distributed to the analysis team.

Validation of intra-exonic tandem repeat expansion in PAX5 exon 5

To validate the intra-exonic tandem repeat expansion in PAX5 exon 5 in case SJALL068100_D1 (PAUKGC), RNA was reverse transcribed to cDNA by SuperScript™ III First-Strand Synthesis System (catalogue number 18080051, ThermoFisher Scientific) and amplified by using the KAPA 2GFast HotStart Ready Mix (catalogue number KK5608, Kapa Biosystems) and the following primer pair: forward primer – 5'-GACAGGACATGGAGGAGTGAATC-3' and reverse primer – 5'-ATAGGTGCCATCAGTGTTTGGTG-3'. Amplicons products were checked on agarose gel, purified by Wizard® SV Gel and PCR Clean-Up Systems (catalogue number A9282, Promega) and cloned into Zero Blunt™ TOPO™ vector (catalogue number 45-1245, ThermoFisher Scientific). The next day colonies were picked and subjected to miniprep culture, plasmid isolation by Wizard® Plus SV Minipreps DNA Purification Systems (catalogue number A1460, Promega) and Sanger sequencing. Sequencing data were analyzed by CLC Main Workbench (Qiagen).

Nanopore sequencing

PCR amplicons were prepared for Oxford Nanopore sequencing using the ligation sequencing kit with native barcoding (SQK-NBD112.24) per the manufacturer's instructions and sequenced for 24 hours on an R10.4 (FLO-MIN112) flow cell. Basecalling was performed in real time through MinKNOW using Guppy v6.3.8 in "super-accurate" mode. Reads were aligned to the PAX5 CDS reference using Minimap2 to identify isoforms.[10]

Genomic analysis summary

An ensemble approach was applied to identify somatic mutations (SNV/indels) with 5 variant callers. The consensus call sets were annotated and manually reviewed. Variants with weak support were excluded. Somatic copy number alternations (SCNA) were determined via CONSERTING.[11] Somatic structural variants (SV) were identified though five structural variant callers. A union set of SV was generated and annotated followed by manual review for alignment quality and read supports. Mutational signatures were profiled by fitting SNV and indel counts per 96 tri-nucleotide contexts to the COSMIC signatures version 3.2. Signatures with <5% overall contribution were excluded from the summary.

Whole genome and exome sequencing mapping and variant calling

The paired end sequencing WGS and WES reads were mapped with BWA-MEM[12] to human Hg38. The alignment quality was assessed using Qualimap.[13] An ensemble approach was applied to identify somatic mutations (SNV/indels) with multiple published tools, including Mutect2 (v4.1.2.0),[14] SomaticSniper (v1.0.5.0),[15] VarScan2 (v2.4.3),[16] MuSE (v1.0rc)[17] and Strelka2 (v2.9.10).[18] The consensus calls by at least two callers were considered as confident mutations. The consensus call sets were further manually reviewed for the read depth, mapping quality, and strand bias to remove additional artifacts. Variants called by a single caller were rescued subsequently after variant quality review. Annovar[19] was used for variant annotation.

Somatic copy number alternations (SCNA) were determined via CONSERTING.[11] For somatic structural variants, five SV callers were used to generate a union set of SV events, including Delly (v0.8.2),[20] Manta (v1.5.0),[21] and GRIDSS (v2.5.0),[22] LUMPY (v0.3.0)[23] and novoBreak (v1.1.3rc).[24] The SV calls passing the default quality filters of each caller were merged into a union set of SVs using SURVIVOR (v1.0.7)[25] and genotyped by SVtyper (v0.7.1).[26] The intersected call sets were manually reviewed for the supporting soft-clipped and discordant read counts at both ends of a putative SV site.

The deletion and amplification of specific genes, including *IKZF1*, *CDKN2A*, *CDKN2B*, *PAX5* and *ERG*, were manually reviewed to confirm their status of copy number changes in each sample by combining the WGS coverage profiles, CNV and SV calling results. A cutoff of 20Mb was used to define focal events.

Mutation burden and mutation signature analyses

The mutation burden was quantified for all samples sequenced by WGS and WES. All SNV and indels passed QC and review were analyzed. The genome length of Hg38 exonic regions was 87,343,287 base pairs. The total number of exonic SNVs and indels per sample was divided by this number and multiplied by 10^6 to obtain the number of SNVs per Mb for each sample. Mutational signatures were profiled by fitting SNV and indel counts per 96 tri-nucleotide contexts to the COSMIC signatures version 3.2 using MutationalPatterns.[27] Signatures with <5% overall contribution were excluded from the summary plot.

RNAseq analyses

The adapters in sequencing reads were trimmed with Trim Galore (v0.4.4; https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, -q 20 –phred 33 --paired). The trimmed sequencing reads were mapped with STAR[28] to human genome GRCh38. The expected gene counts calculated using RSEM for each sample were compiled to one gene count matrix. Only genes annotated as level 1 or 2 by GENCODE (v31) were kept in the downstream analysis. In addition, only genes with count per million (CPM) more than 0.5 in at least one sample were kept. The normalization factor for each sample was calculated using “calcNormFactors” in the “edgeR” package (v3.26.8),[29] and gene expression values were transformed and normalized using voom[30] in the “limma” package (v3.40.6)[31] in R.

Sample subtyping

Subtyping of MP2PRT B-ALL samples followed the procedure as described previously[32,33] based on cytogenetics, WGS copy number profiles, gene expression, and fusions or rearrangements

detected by RNA-seq and/or WGS. Prediction Analysis of Microarrays (PAM)[34] was used to identify subgroups with distinct gene expression profiles.[35] Samples without RNA-seq were usually assigned to the B-other subtype; however, in cases without RNA-seq but with WGS, subtype-defining rearrangements detected by WGS were used to assign subtypes. Cytogenetic and FISH data collected on the clinical trials was compared against the results of genomic analyses, and discrepancies were resolved by in depth review.

Subclassification of PAX5alt and IKZF1 cases

Sub-Classification of PAX5alt

The sub-classes of PAX5alt was defined based on the types of lesions present in the PAX5 genes in the order of amplification of PAX5 (PAX5amp), PAX5 gene fusions (Fusion), small SNV/indels (Mutation), other structural variants (SV) and *PAX5::JAK2* fusion (*PAX5::JAK2*). A more granular sub-classification was defined by the co-current types of PAX5 lesions (Fig. 3).

Classification of IKZF1^{plus}

IKZF1^{plus} status is commonly examined by targeted DNA copy number alteration profiling using multiplex ligation-dependent probe amplification arrays (MLPA) and is defined as the presence of *IKZF1* deletion with concomitant *CDKN2A/CDKN2B* homozygous deletion, and/or *PAX5* deletion, and/or deletion of pseudoautosomal region 1 (PAR1) of chromosome Xp/Yp, without deletion of ERG.[36] In this definition, PAR1 deletion is used a surrogate marker of rearrangement of *CRLF2*, as it accompanies formation of *P2RY8::CRLF2*; and *ERG* deletion is used as a surrogate marker of *DUX4* rearrangement. However, neither PAR1 nor *ERG* deletion detect all cases of *CRLF2* or *DUX4* rearrangement, respectively. Moreover, the MLPA assay cannot distinguish focal *IKZF1* alteration from broad loss of chromosome 7 or 7p. Thus, we used genomic data to simulate a MLPA-based definition, and also utilized a genomically-faithful definition of IKZF1^{plus} that considers all *CRLF2* rearrangements, and all *DUX4* rearrangements, and focal (<20Mb) deletions of *IKZF1*.

Classification and regression tree (CART) analysis

The Hyperdiploid samples were split with 80% of relapse and non-relapse samples (n=410) as training set and the remaining 20% (n=102) as testing datasets. We use R implementation of CART to build the CNV based decision tree.[37] For training, the following parameters were used: max allowed depth of 30 (max-depth constraint by rpart), the minimum number of observations that must exist in a node to split (minsplit) as 20 and the complexity parameter ("cp") as 0.01. The repeat k-fold cross-validation (k=10, repeat 5 times) of Caret[38] was used to get list of pruning parameters and choose best pruning parameter. For statistical associations, we used Fisher's exact test to compare samples from one node and the rest. Survival analysis was also used for visualization and the log-rank test used to show the significance of the decision rules.

Variant annotations, transformations and aggregation

For simplicity, somatic mutations were categorized into eight types: "gain" (copy number gain), "hetdel" (copy deletion), "homodel" (two copy deletion), "SNV" (all non-synonymous affecting protein coding sequence), "indel" (small indels affecting protein coding sequence), "TRA" (chromosomal rearrangements), "INV" (inversion) and "INS" (large insertions/duplications). Annotation data of 60,568 protein coding genes and non-protein coding processed transcripts were retrieved from Ensembl BioMart database (release 104) using 'biomaRt' R package. Thereafter, Genomic Random Interval (GRIN) package[39] was used to map each genomic lesion to the list of annotated genes based on their genomic coordinates (start and end positions) to identify genomic loci affected by each type of genomic lesions. Large chromosome-level or arm-level CNVs were filtered out to avoid mixing of focal CNVs and large CNVs in one analysis. Specifically, CNVs with segment length >80% of the arm size or >10 Mb were filtered. Chromosome-level CNVs in the Hypodiploid subtype were analyzed separately such as in the CART analysis.

Single mutation type-based analysis

For each gene, we considered each mutation type individually. Different types of mutations include SNV, short indel (INDEL), heterozygous deletion (HETDEL), homozygous deletion (HOMDEL), copy number gain (GAIN), translocation (TRA). Mutations were coded as a binary variable indicating whether there is at least one mutation aligned to that gene. We treated the relapse status as a binary trait: 0 if the relapse event is not observed, 1 if the relapse event is observed. Then we used Firth logistic regression to test the association between each gene-based mutation with the relapse status. For the whole sample group analysis, we performed two schemes of association tests. The first analysis scheme was the association test without adjusting for subtypes. This analysis provided the raw or marginal associations of each somatic mutations with the phenotype when subtypes were not considered. The second analysis scheme was the association test including the subtypes as covariates. This analysis provided the conditional associations after the subtype effects were accounted for. Subtypes with sample size <10 were merged into one single subtype category in the second analysis scheme. Our main analysis results were based on treating the relapse status as a binary trait. False positive rates (FDR) were calculated within each sample group using a permutation based approach. Specifically, we permuted the phenotype 1000 times and recalculate the association tests in each permutation. Then these permutation based p-values were used to compute a permutation based FDR,[40] using functions implemented in CoCoRV.[41]

Multiple mutation types-based analysis

In this analysis, we combined different mutation types within a gene together using a dominant model. The combined mutation was coded as a binary variable: 0 when there was no single mutation event in the gene, 1 when there was at least one mutation event in the gene. Once the genotype matrix was generated, all other analyses were the same as in the single mutation type-based analysis.

Association analysis of MRD as the secondary trait

Because the samples were collected as a case-control design based on the relapse status, association analysis between the secondary trait MRD and the genotypes should account for the case-control sampling design. We used the IPW method[42] to account for the case-control design. The IPW method required an estimation of the prevalence of the relapsed cases in the B-ALL sample group. The prevalence was set to 0.1 because for standard risk, patients on AALL0331 (the majority of the patients in this sample group) had a 6-year disease-free survival (DFS) around 89% and the survival curve plateaued at about 5 years.[43] Then weights were calculated for each sample based on the prevalence and the case-control ratio in the full sample group. Weighted Firth logistic regression was then used for the association test. Similar as the analyses for the relapse status, we performed single mutation type-based analysis and multiple mutation combined analysis. For the whole sample group, two analyses were performed, the first was without adjustment for subtypes and the second adjusted for subtypes.

Survival analysis

For most analyses, we treated the samples with the relapse events as cases and the samples without relapse or censored as controls, and used the logistic regression model or Fisher's exact test to perform association tests. All confidence intervals (CIs) reported were at the 95% confidence interval. To account for the sample selection probabilities, we adjusted the Kaplan-Meier curves with inverse probability weights. Specifically, we first calculated the sampling rates of cases and controls. We compared the full sample groups of standard risk patients with matched MRD and standard risk patients in the MP2PRT dataset. The estimated sampling rate is 0.229 for controls and 0.442 for cases. Using inverse sampling probabilities as weights, the weights are 4.37 for controls and 2.26 for cases; equivalent weight 1 for cases and 1.93 for controls. In addition, we used the weighted cox-regression tests to derive the P values. All CIs reported are at the 95th confidence level.

SUPPLEMENTARY RESULTS

Association between mutational signatures and risk of relapse

We performed mutation signature analyses using both single nucleotide variants (SNV) and indels. In the *ZNF384*-rearranged subtype, the median relative contribution of SNV signature SBS89 was higher in the relapse than no-relapse samples (P <0.05, Wilcoxon rank-sum test). SBS89 is potentially most active in childhood tumors without known etiology. Finally, with respect to the indel mutation signature, in the hyperdiploid subtype, the median relative contribution of signature ID2 was higher in no-relapse (P=0.014, Wilcoxon rank-sum test), while ID1 was higher in relapse (P=0.019, Wilcoxon rank-sum test). In *ZNF384*-rearranged ALL, the relative contribution of ID2 is higher in the relapse samples (P=0.047). ID1 and ID2 signatures are related to DNA replication slippage and found frequently in cancers with DNA mismatch repair deficiency.

Unbalanced translocations and the risk of relapse

In CCR patients, the unbalanced version of the t(1;19) translocation (i.e. der(19)t(1;19)) was present in 28/37 (75.7%) patients, whereas relapse patients showed enrichment of the balanced t(1;19) (11/17 (64.7%)) versus the unbalanced der(19)t(1;19) (6/17 (35.5%)) (P=0.004, χ-square test). Early studies indicated that there was a poorer outcome for patients with the balanced t(1;19),[44,45] with subsequent studies showing no difference in outcome for the balanced versus the unbalanced form of the translocation.[46-48] The observations in this study were consistent with the earlier studies. Of note, in all patients with the unbalanced version of the t(1;19), chromosome 1 retained heterozygosity. These observations support the model proposed by Paulsson et al.[49] that an unbalanced t(1;19) most likely arises from an initial trisomy of chromosome 1 followed by a translocation (1;19) and the subsequent loss of the derivative chromosome 1. Indeed, gain of a number of genes on chromosome 1 was strongly associated with a reduced risk of relapse (OR=0.12, p= 0.00085) for patients harboring *TCF3::PBX1*.

Association analysis of MRD as the secondary trait

Within common subtypes, the majority of 810 gene-level associations with MRD (FDR <0.2, Table S14b) were observed in the *PAX5*alt group, most of which (702 of 797 genes) were due to deletions of chromosomes 9 (OR=0.18 CI=0.05-0.52, *P*=0.001) and 20 (OR=0.22, CI=0.06-0.65, *P*=0.006, Table S15) that accompany the formation of structural variants of *PAX5*, such as unbalanced rearrangements.

SUPPLEMENTARY TABLES

The following tables are in the Supplementary Table Excel workbook:

Supplementary Table 1. Initially selected cases and controls from AALL0331, AALL0932 (NCI SR) and AALL0232, AALL1131 (NCI HR)

Supplementary Table 2. Final analyzed cases and controls in Table S1 with adequate sequencing results

Supplementary Table 3. Clinical differences between study group those selected but not analyzed

Supplementary Table 4. Clinical and sequencing information study group

Supplementary Table 5. List of coding-region somatic SNVs and indels

Supplementary Table 6. List of somatic structural variants from WGS and WTS

Supplementary Table 7. List of somatic DNA copy number alterations from WGS

Supplementary Table 8. Matrix of SNVs, indels, CNVs and SVs tabulated by cases

Supplementary Table 9. Genetic subtypes identified in the study group

Supplementary Table 10. Association analysis of clinical factors with relapse and MRD status

Supplementary Table 11. Mutational signatures of SNV and indel mutations

Supplementary Table 12. List of PAX5 alterations in the PAX5alt group.

Supplementary Table 13. Mutation frequencies of putative leukemia driver genes
Supplementary Table 14. Statistical result of genetic alteration association analysis

Supplementary Table 15. Representative genome-wide associations with phenotypes driven by large deletions.

Supplementary Table 16. Significant association results (FDR <0.2) between putative leukemia driver genes and relapse or MRD for individual mutation types.

Supplementary Table 17. Significant association results (FDR < 0.2) between putative leukemia driver genes and relapse or MRD for combined mutation types.

Supplementary Table 18. Association results between IKZF1 alterations and the relapse status for individual subtypes.

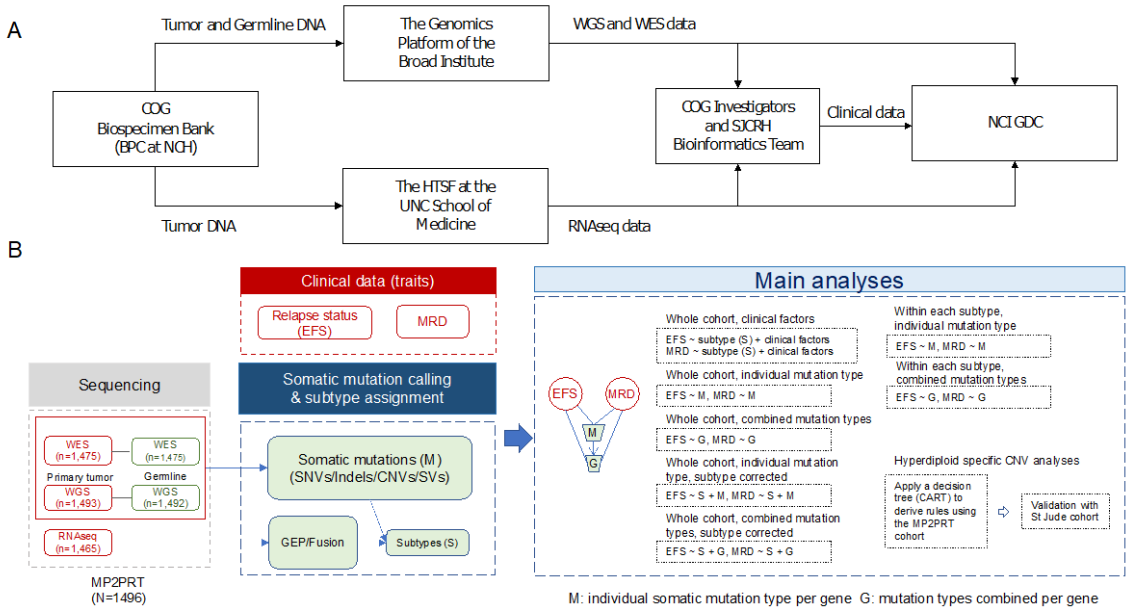
Supplementary Table 19. List of IKZF1 alterations

Supplementary Table 20. Case number of IKZF1plus

Supplementary Table 21. Cochran–Mantel–Haenszel test (CMH) results stratified by the groups defined using different CNV based rules.

Supplementary Table 22. Table of sample counts within each cluster defined from CART

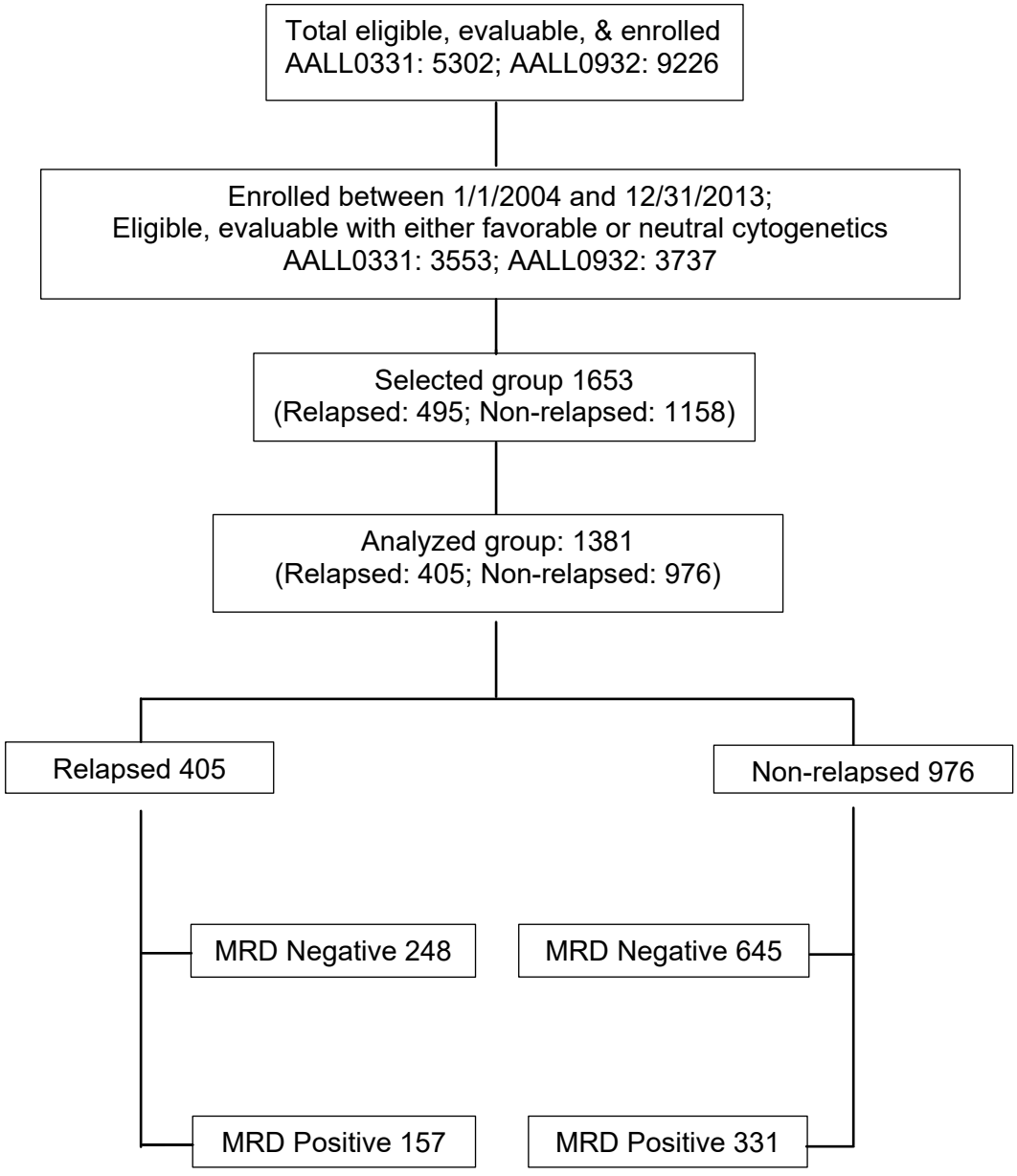
SUPPLEMENTARY FIGURES



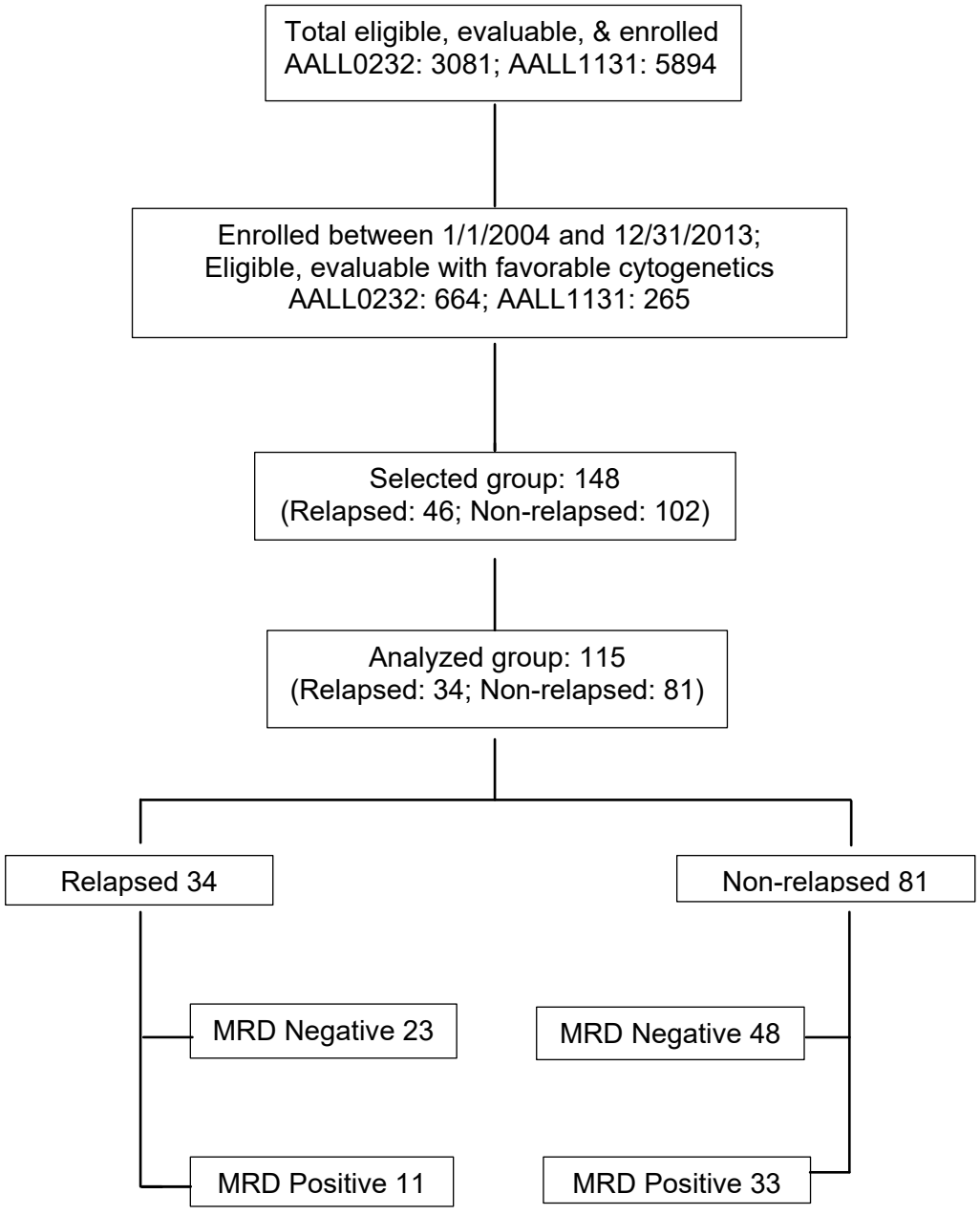
Supplementary Figure 1. Workflow of specimen processing and analysis.

A, Biospecimen processing, genomic characterization and genomic data analysis were conducted under the NCI Cancer Moonshot℠ Molecular Profiling to Predict Response to Treatment (MP2PRT) Program. The COG Biospecimen Bank, located within the Biopathology Center (BPC) at Nationwide Children's Hospital (NCH), performed nucleic acid extractions and shipped qualified tumor DNA, matching germline DNA and tumor RNA to two genomic sequencing centers. The Genomics Platform of the Broad Institute of MIT and Harvard performed whole genome sequencing (WGS) and whole exome sequencing (WES), and the High-Throughput Sequencing Facility (HTSF) at the University of North Carolina (UNC) School of Medicine performed total RNA sequencing (RNAseq). Sequencing data was provided to the COG Investigators and a bioinformatics team at the St. Jude Children's Research Hospital (SJCRH) for genomic data analysis and to the NCI Genomic Data Commons (GDC) for data sharing purposes. Relevant clinical data has also been made available at the NCI GDC. **B**, the workflow of the data collection, genomics analyses and association analyses. GEP: Gene expression profile.

A **NCI Standard Risk**

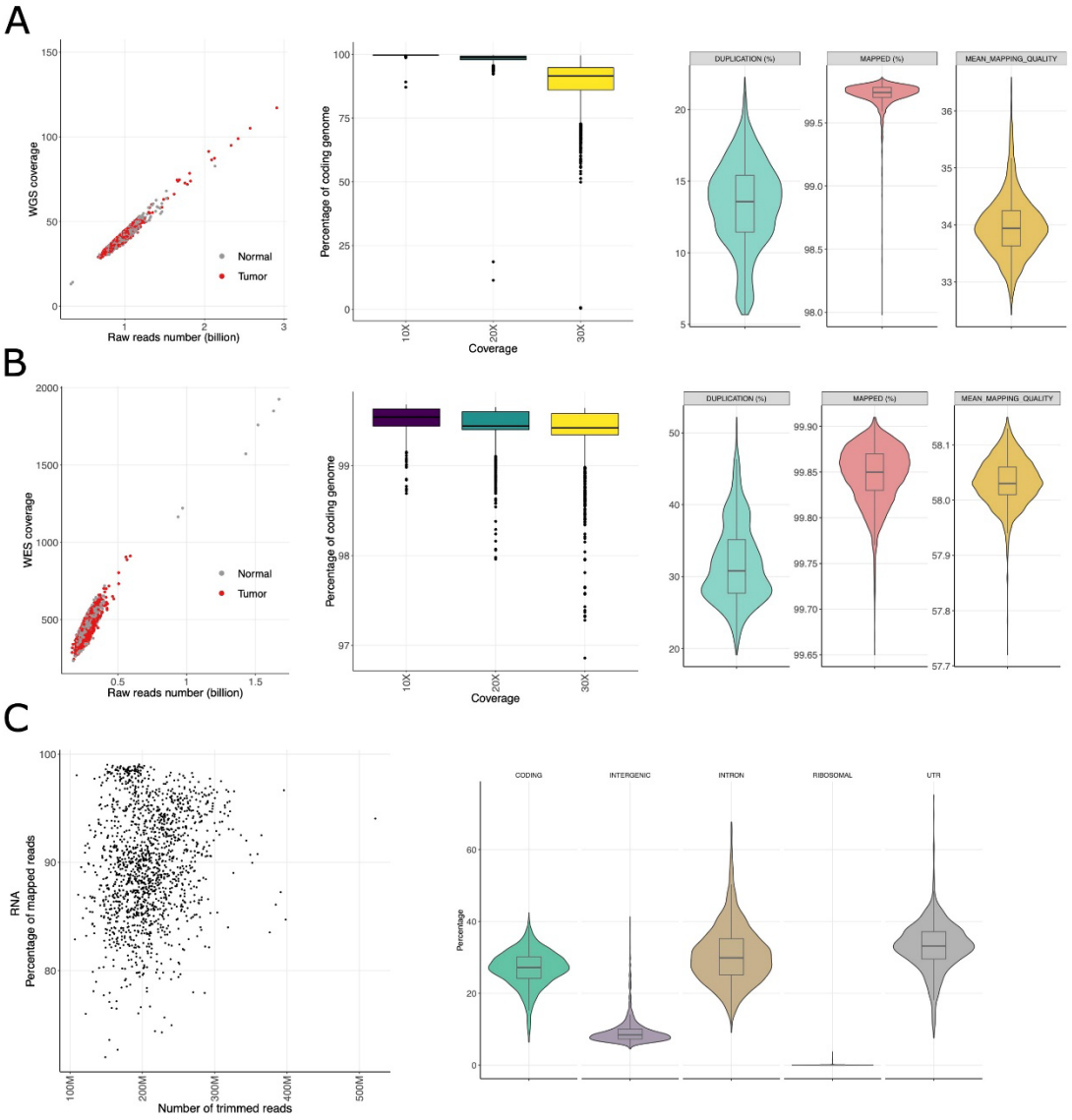


B **NCI High Risk**



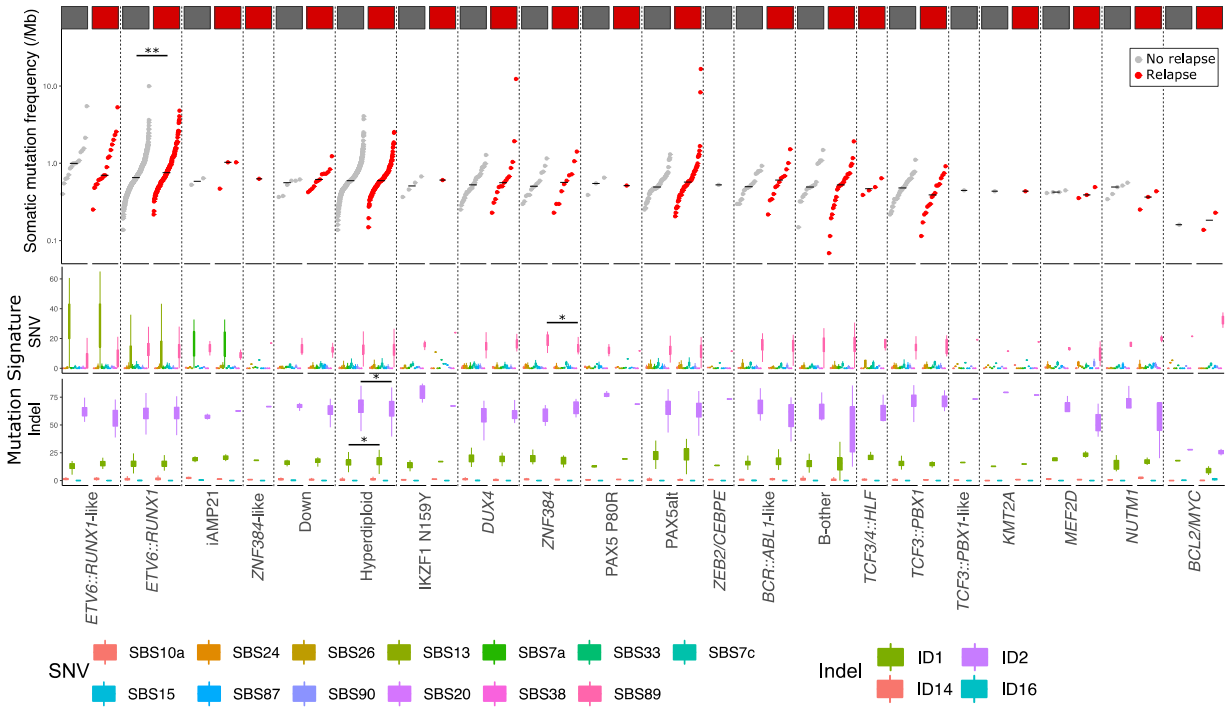
Supplementary Figure 2. Consort Diagram of selected versus analyzed samples, including those who relapsed versus those who did not, stratified by MRD < .01 vs ≥ 0.01%.

A. Selected NCI SR cases and controls from AALL0331 and AALL0932 studies; **B**. Selected NCI HR cases from AALL0232 and AALL1131 studies.



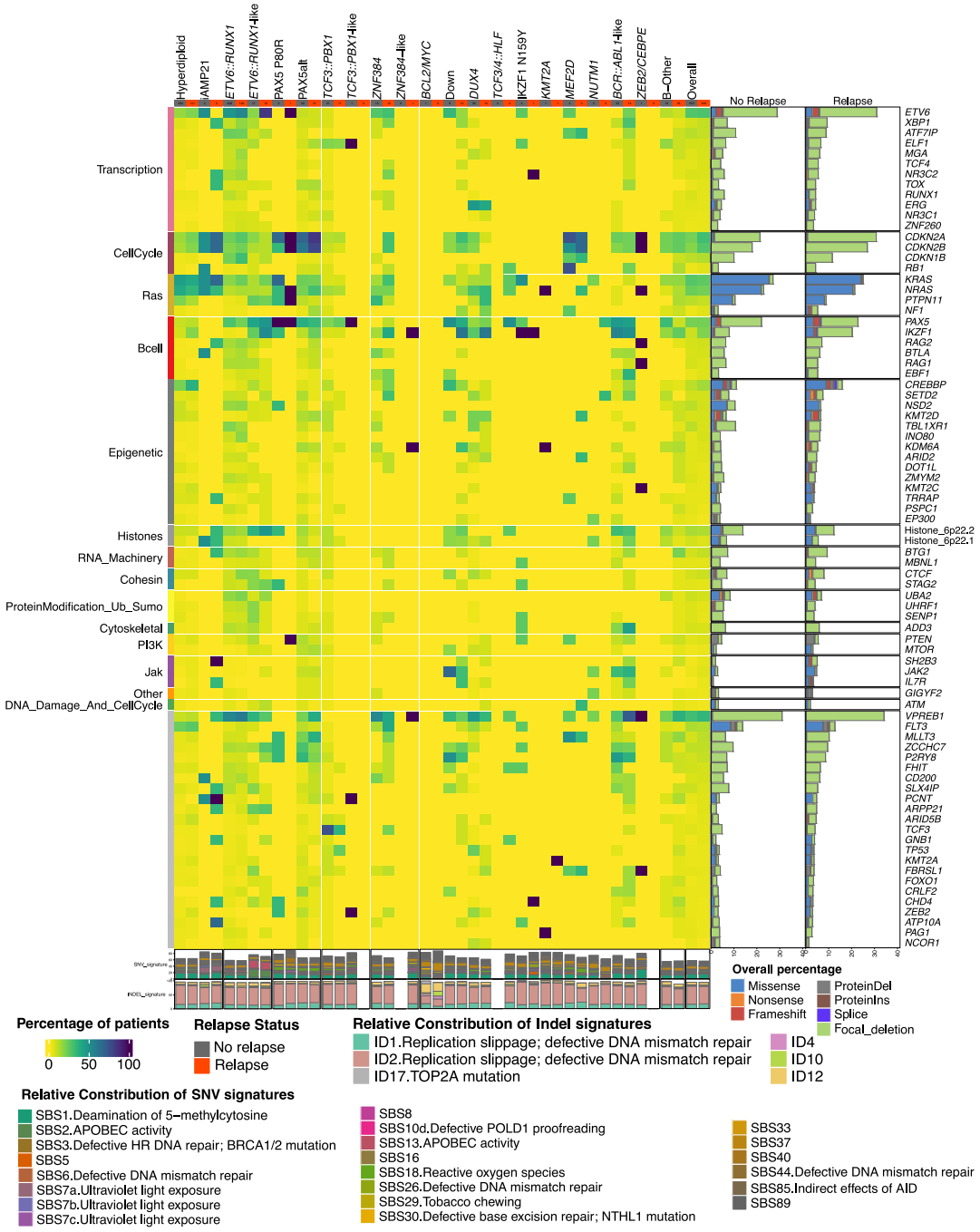
Supplementary Figure 3. Sequencing data quality of MP2PRT.

A, Sequencing quality of whole-genome sequencing (WGS) data (n=1,493). Left, correlation of WGS coverage and raw read number. Middle, boxplots showing the distribution of percentage of genome per sample with at least 10X, 20X, and 30X fold coverage. Right, violin and boxplots showing the distribution of percentage of duplicated reads, percentage of mapped reads, and mean mapping quality across samples. **B,** Sequencing quality of whole-exome sequencing (WES) data (n=1,475). Left, correlation of WES coverage and raw read number. Middle, boxplots showing the distribution of percentage of genome with at least 10X, 20X, and 30X fold coverage. Right, violin and boxplots showing the distribution of percentage of duplicated reads, percentage of mapped reads, and mean mapping quality across samples. **C**, Sequencing quality of RNAseq data (n=1,465). Left, positive correlation of number of trimmed reads and number of reads mapped is shown. Right, percentage of RNAseq reads mapped in coding, intergenic, intronic, ribosomal and UTR regions.



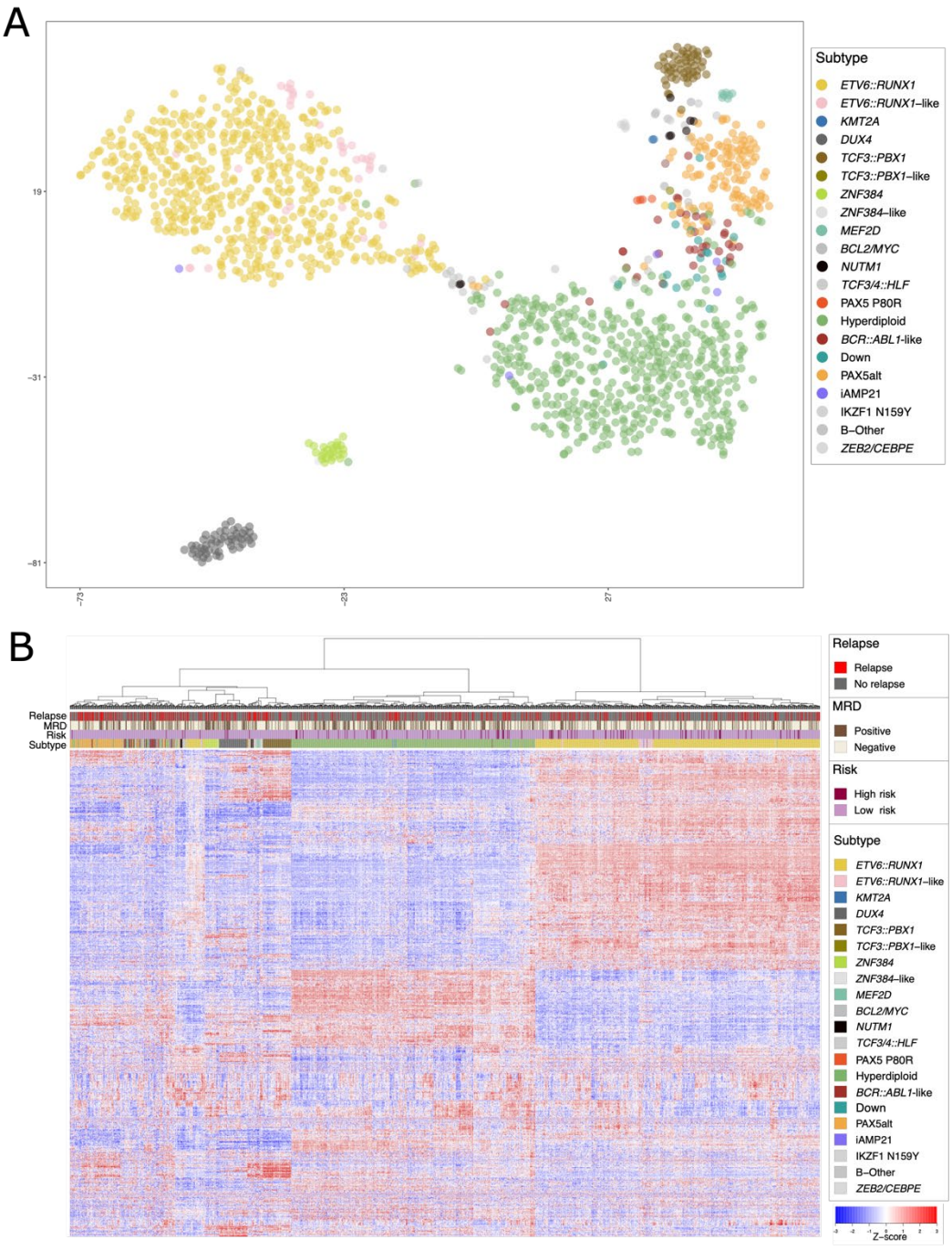
Supplementary Figure 4. Somatic mutation frequency across all subtypes.

The median somatic mutation frequency is shown as a black bar. Note that the highest somatic mutation burden occurred in *ETV6::RUNX1* and Hyperdiploidy, with significant differences between cases and controls in *ETV6::RUNX1*. Middle, relative ratio of SNV signatures with significant differences between relapse and no-relapse samples in respective comparisons (*P*<0.05, Wilcoxon rank-sum test). Differences were significant for *ZNF384* cases and controls. The comparisons of signatures with average relative contribution >10% showing significant difference between relapse and no-relapse are annotated with asterisk (*, *P*<0.05; **, *P*<0.01). Bottom, relative contribution of indel signatures with noted significant differences between case and control samples in respective comparisons, notably in hyperdiploid and *ZNF384* subtypes (*P* < 0.05, Wilcoxon rank-sum test). The comparisons of signatures with average relative contribution > 10% showing significant difference between relapse and no-relapse are annotated with asterisk (*, *P*<0.05; **, *P*<0.01).



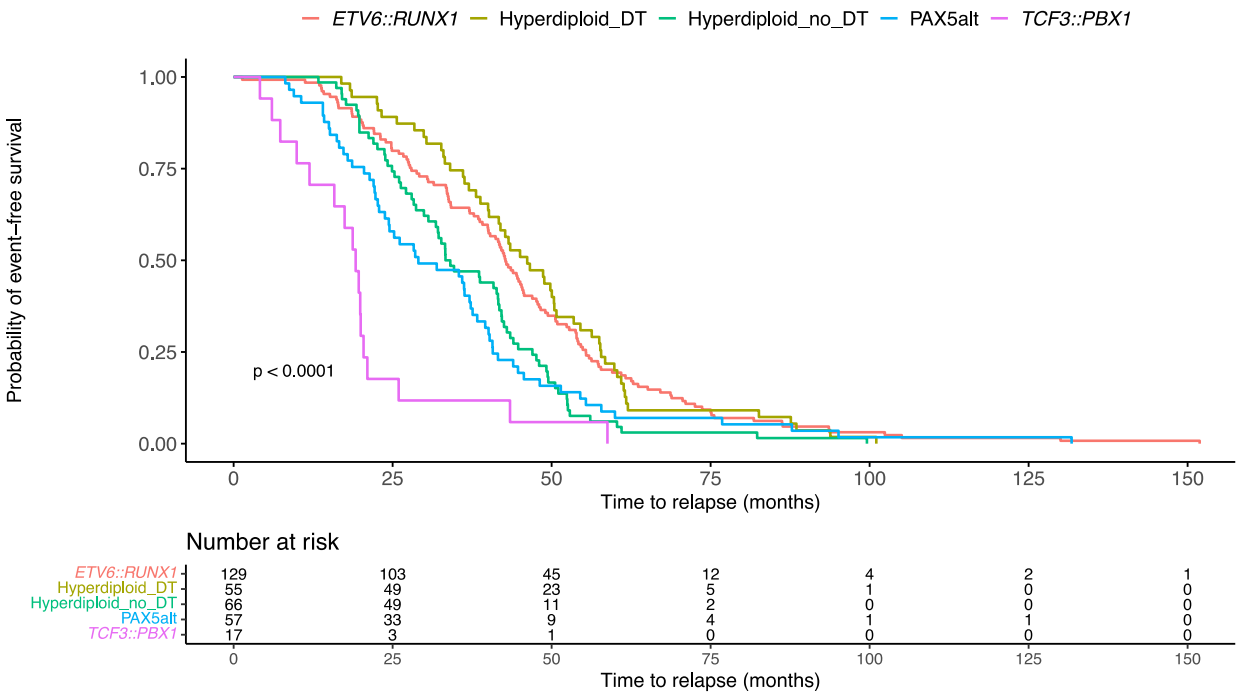
Supplementary Figure 5. Mutational landscape of the study group.

The genomic characterization of somatic alternations in 1,496 samples. The heatmap (left) shows the percentage of relapse (case) or no relapse (control) samples of each subtype (paired columns) with somatic SNV/indel alterations and/or focal deletion in the driver genes (rows). Only variants with overall percentage >3% are shown. Subtypes and sample numbers are annotated in the top. The histograms (stacked bars, middle) shows the percentage of samples with alterations in each gene in the case control groups, colored by the alteration types. Bottom, relative contribution of mutation signatures with percentage >5% in case control samples in at least one subtype is shown.



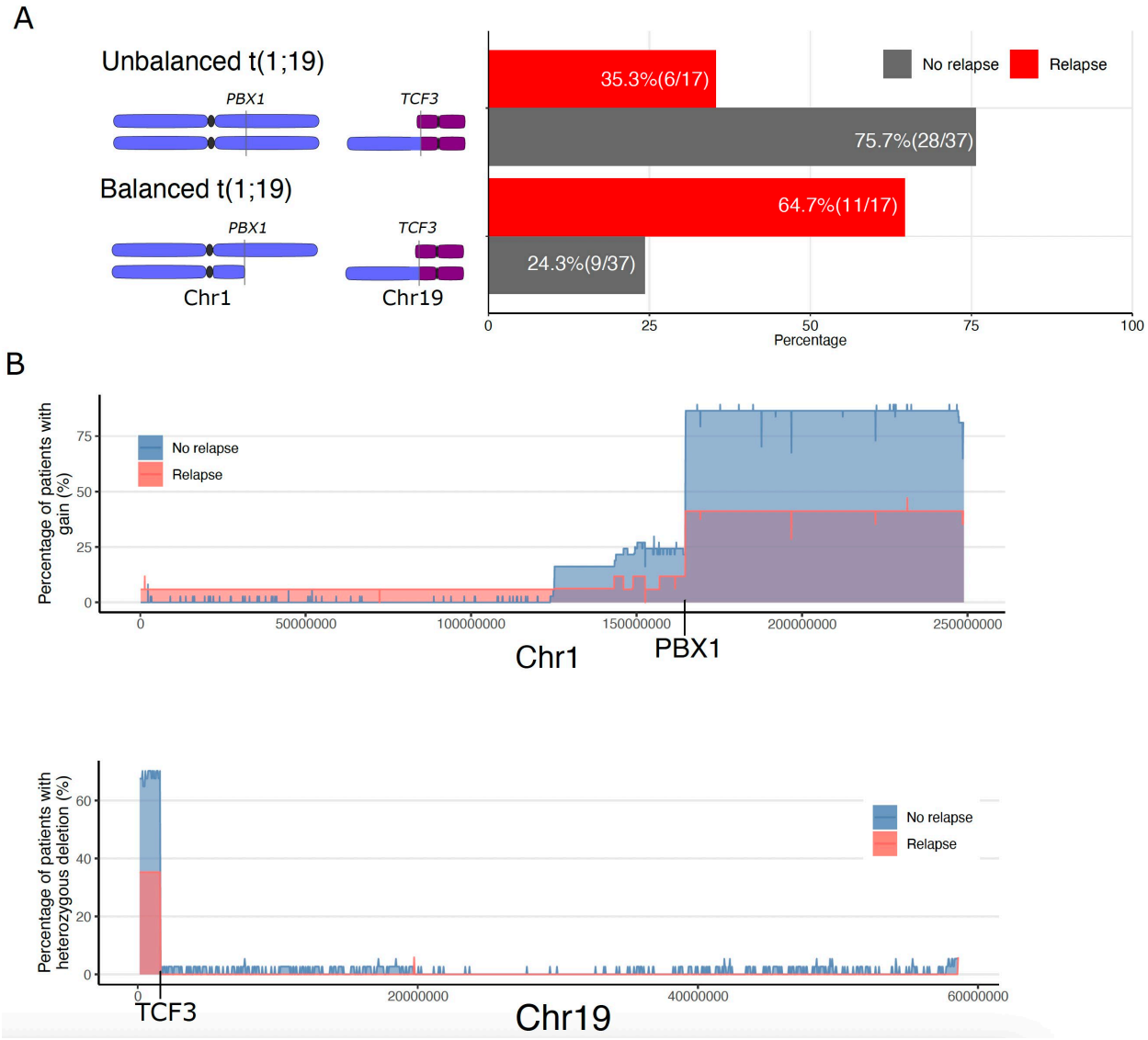
Supplementary Figure 6. Transcriptome clustering of MP2PRT samples by driver genomic subsets.

A, tSNE analysis of 1,465 samples with RNA-seq. Each solid point represents one MP2PRT sample. The samples are colored by subtypes as indicated in the legend. **B,** Unsupervised hierarchical clustering and heatmap showing top differentially expressed genes (n=1,000) in MP2PRT. A total of 1,372 samples with adequate transcriptome is shown; low quality samples with duplicate rate > 70% or coverage of coding regions < 30X were excluded (n=93). The status of relapse, NCI risk stratification and subtypes are annotated at the top.



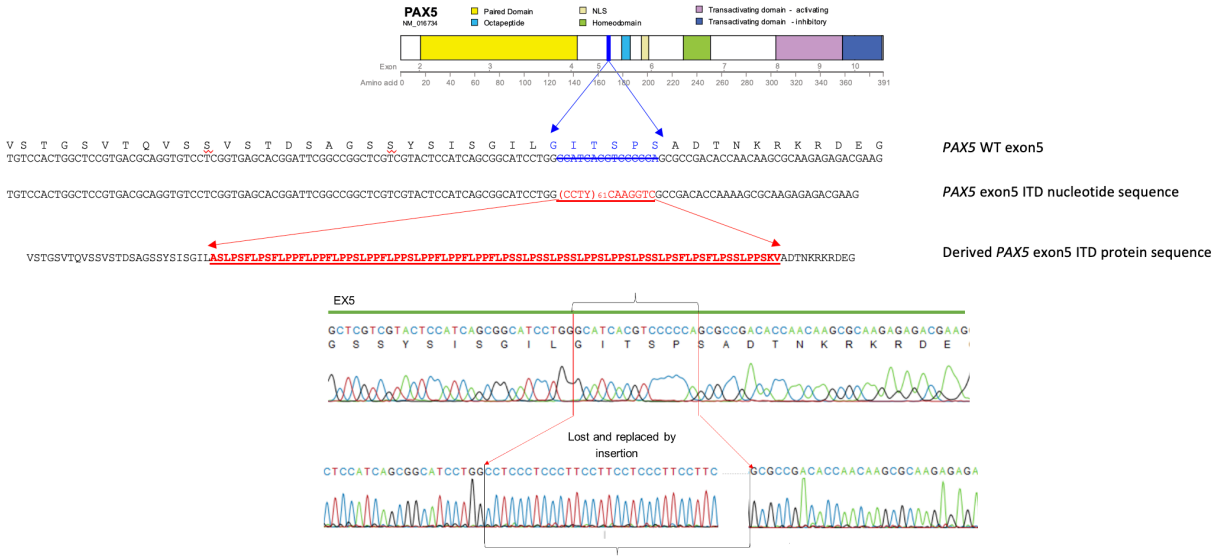
Supplementary Figure 7. Distinct patterns of time to relapse among B-ALL genetic subtypes.

Numbers of subtypes and relapse of the 1,496 analyzed MP2PRT study group with distinct time to relapse patterns of genetic subtypes.



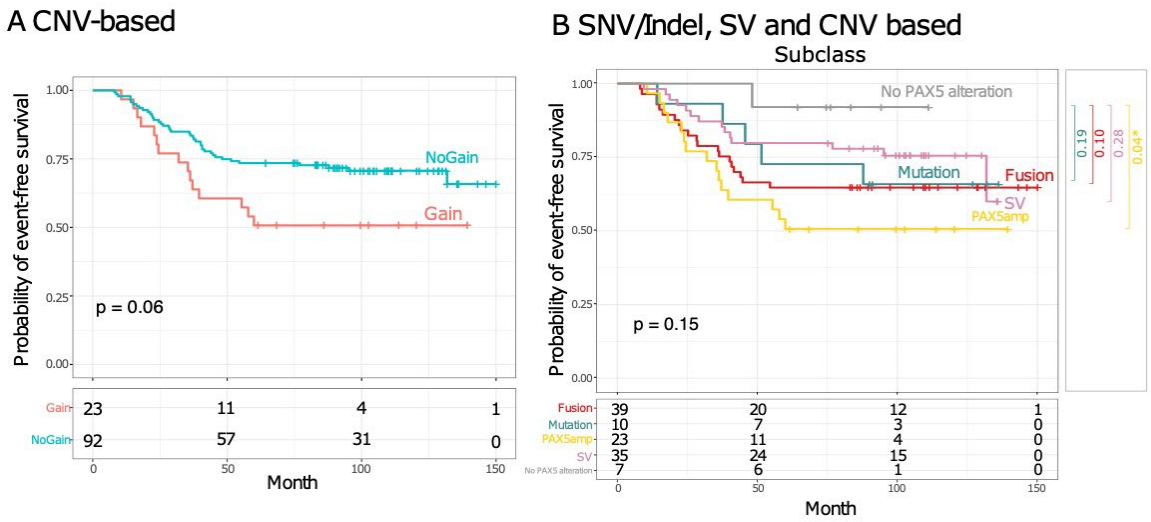
Supplementary Figure 8. Unbalanced translocation and risk of relapse in *TCF3::PBX1* B-ALL.

A, Comparison of percentage of patients with balanced and unbalanced *TCF3::PBX1* translocation. Balanced *TCF3::PBX1* translocation was enriched in cases versus controls. A χ-square test of independence was performed to examine the relation between status of relapse and unbalanced/balanced translocations. The relation between these variables was significant, (*P*=0.0043). Unbalanced *TCF3::PBX1* translocation was more prevalent in non-relapse samples. Comparisons of percentage of patients with gains on chr1 (top), heterozygous deletion on chr19 (bottom) in the *TCF3::PBX1* subtype. The comparisons were performed by sliding window analyses with a bin size of 50kbp.



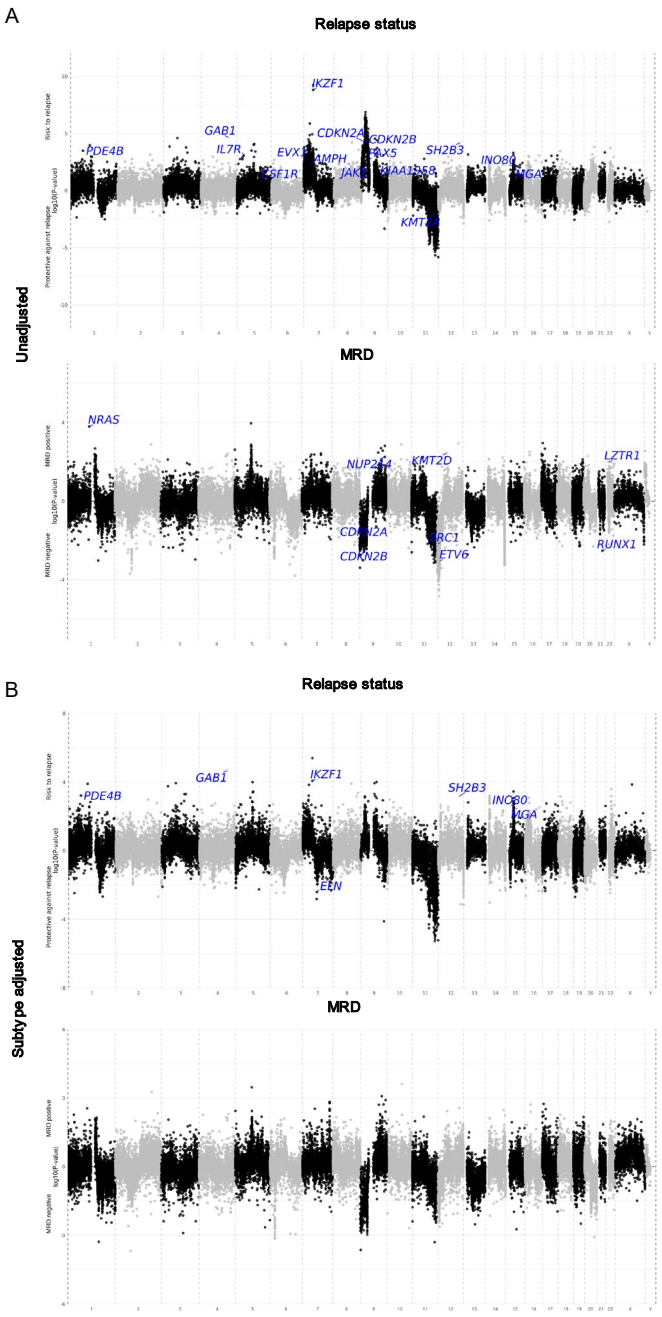
Supplementary Figure 9. Intra-exonic tandem repeat expansion in PAX5 exon 5.

In WGS tumor sample SJALL068100_D1 (PAUKGC), we detected soft-clipping reads supporting a novel ITD event, replacing the original 15bp exonic sequence coding for GITSPS (aa 186-190, NM_016734). The supporting evidence is also observed in the RNAseq of the same tumor. Further PCR cloning and sequencing of the cDNA from the tumor with Oxford nanopore long reads reveals a major resulting clone with 249bp in-frame insertion mainly contributed by 61 repeated tetramer sequences of CCTT and CCTC of variable alterations. The evidence of the tandem repeats from Sanger sequence was also shown, although the exact length of repeats was beyond the detection capability of this assay. The putative derived protein sequence is enriched with Serine (S), Proline (P), Phenylalanine (F), and Leucine (L). The exact function of this insertion is unclear.



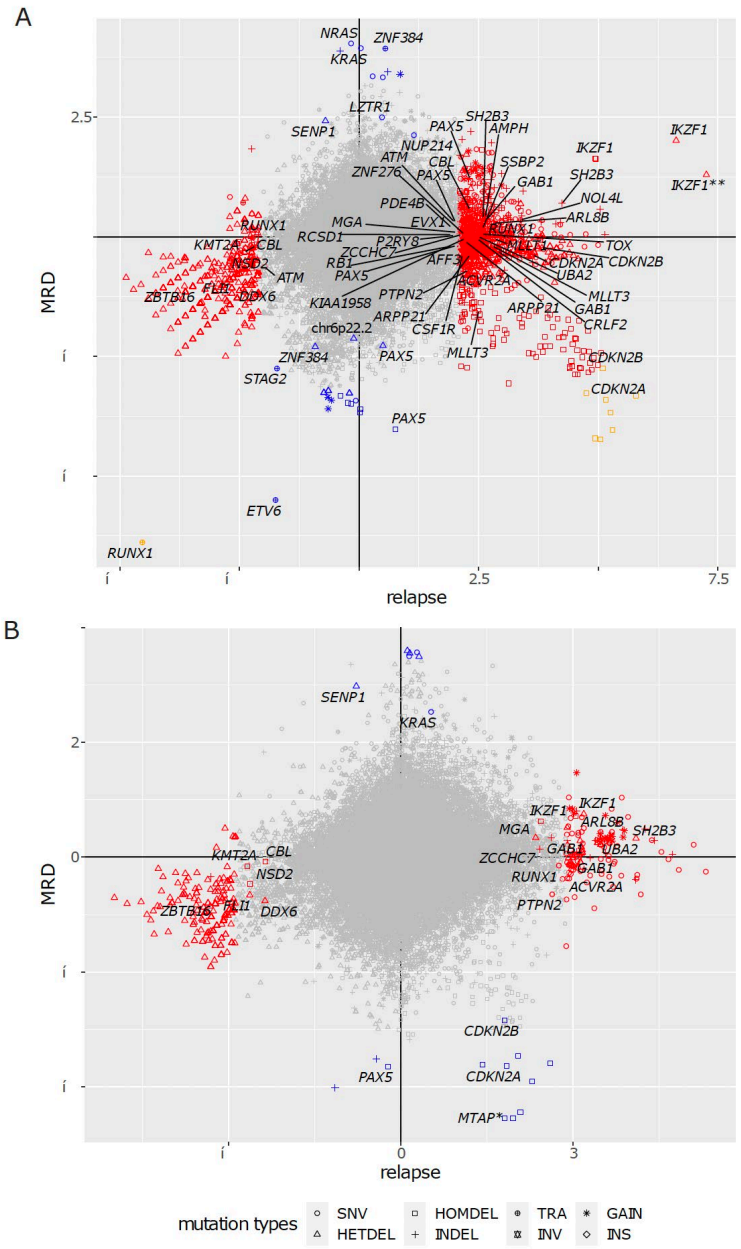
Supplementary Figure 10. Outcome of B-ALL according to PAX5alt lesion

Weighted Kaplan-Meier (KM) curves of subclasses of PAX5 variants in the PAX5alt subtype. A. KM curves for groups classified based on the CNV status. B. KM curves for groups classified based on SNV/indels, SV and CNV. The weighted cox-regression test results are shown on the right when available.



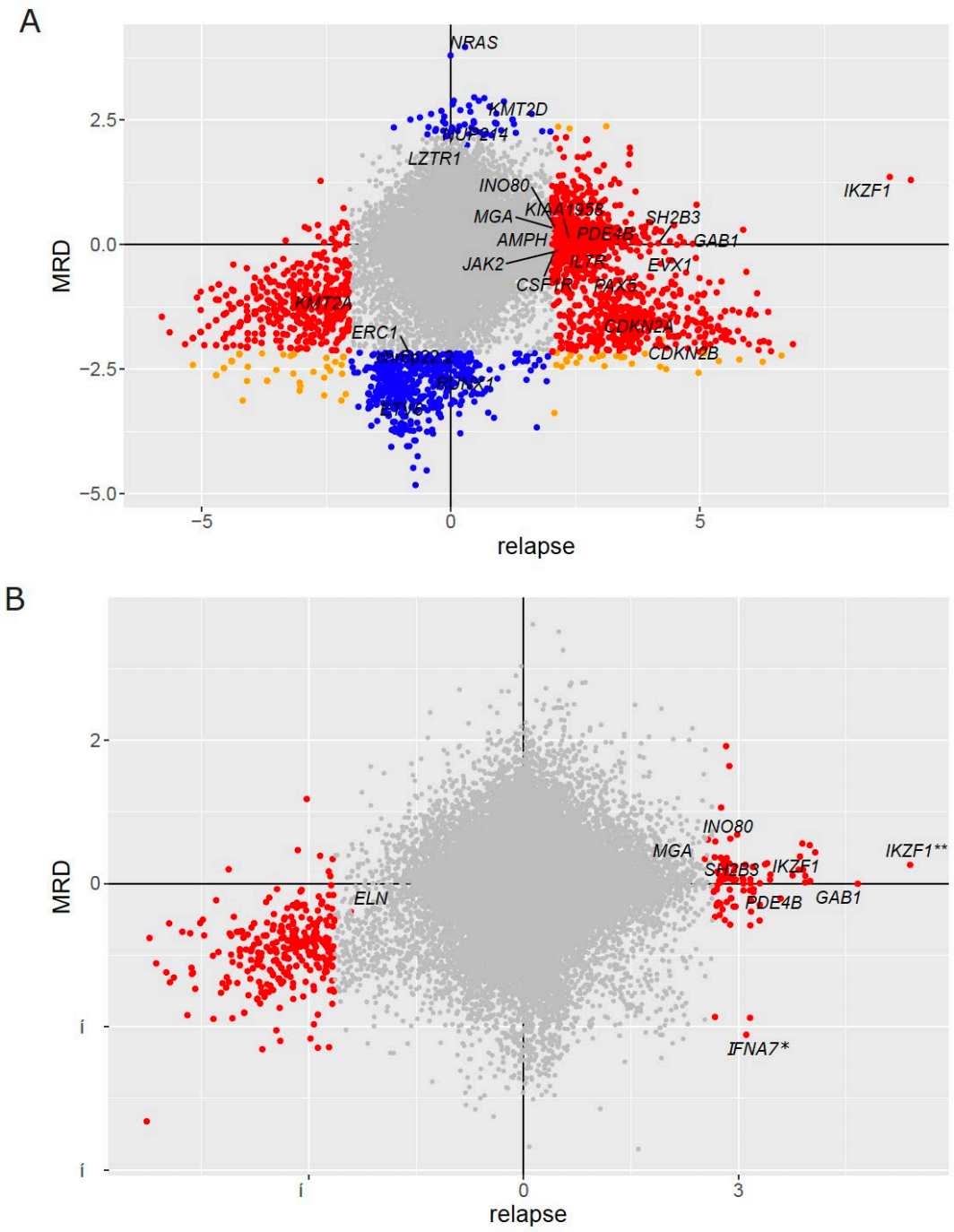
Supplementary Figure 11. Genome wide analysis identifies molecular markers associated with relapse in SR B-ALL.

A. Manhattan plot for relapse status (top) and MRD (bottom) for the entire study group. The analyses were not adjusted for subtypes. **B**. Manhattan plot for relapse status (top) and MRD (bottom) for the entire study group adjusted for subtypes. Associations were performed for each gene combining all alteration types. Genes with FDR <0.2 from the putative leukemia driver gene set based analysis were labeled. For the bottom plot of Panel B, no gene reached FDR <0.2 after adjusting for subtypes for the putative leukemia driver gene set based analysis.



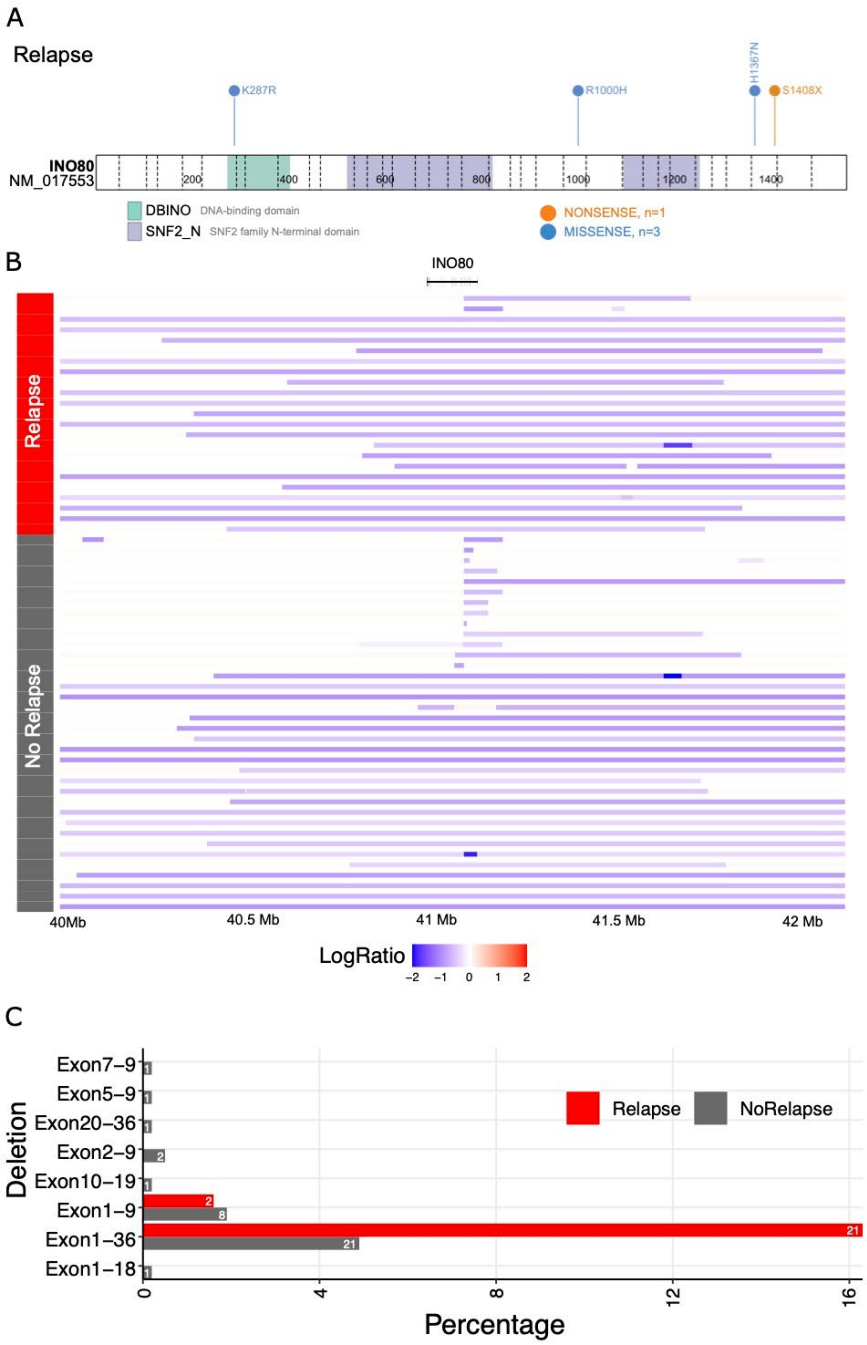
Supplementary Figure 12. Association of single gene alterations with MRD and relapse status.

A, Results without adjusting for subtypes. **B**, Results after adjusting for subtypes. Blue, red, and orange colors indicate FDR < 0.2 (genome-wide or candidate genes-based) for MRD, relapse, or both, respectively. Labelled significant genes are either from the putative leukemia driver genes or manually selected (*). IKZF1**: a subregion of *IKZF1* encompassing exons 7-8 (chr7:50388489-50400097). SNV: single nucleotide variant; HETDEL: heterozygous deletions; HOMDEL: homozygous deletions; INDE: small insertion or deletion; TRA: translocation; INV: inversion; GAIN: copy number gain; INS: insertion.



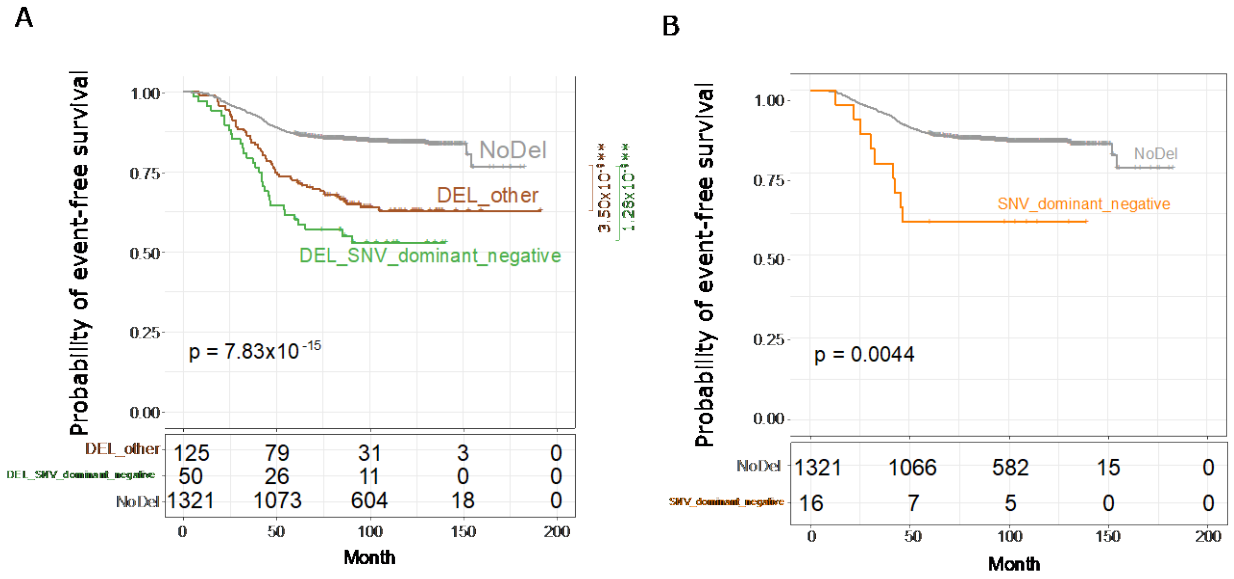
Supplementary Figure 13. Association of multiple alterations within a gene with MRD and relapse status.

A, Results without adjusting for subtypes. **B**, Results after adjusting for subtypes. Blue, red, and orange colors indicate FDR <0.2 (genome-wide or candidate genes-based) for MRD, relapse, or both, respectively. Labelled significant genes are either from the putative leukemia driver genes or manually selected (*). *IKZF1***: a subregion of *IKZF1* captured by ENSG00000285165.



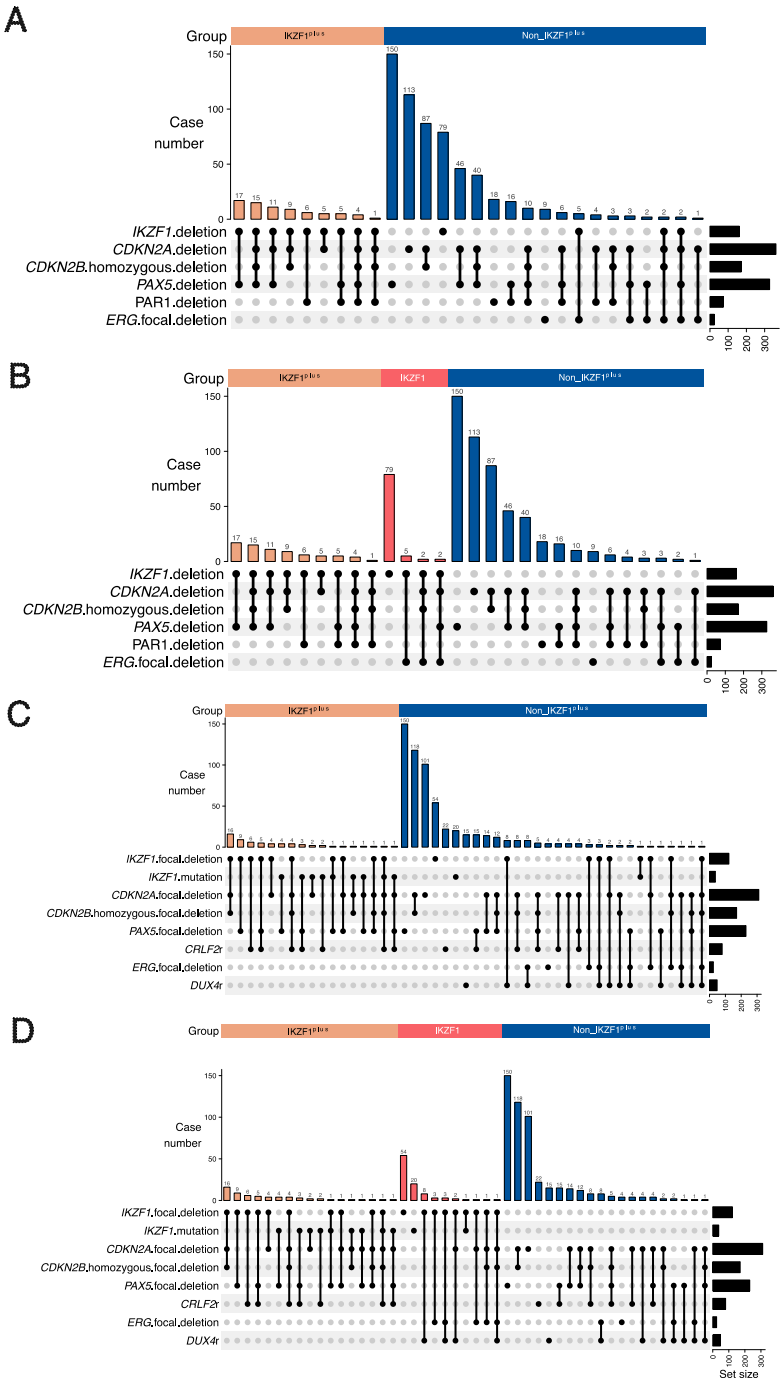
Supplementary Figure 14. Somatic *INO80* deletions are enriched in *ETV6::RUNX1* patients who experienced relapse.

A. Somatic SNVs in *INO80* in *ETV6::RUNX1*. **B**. Segmentation plot of samples with the *INO80* somatic copy number alterations in the *ETV6::RUNX1* subtype. **C**. Percentage of samples with *INO80* deletions. Deletions of the entire locus (exons 1-36), were enriched in the relapse samples.



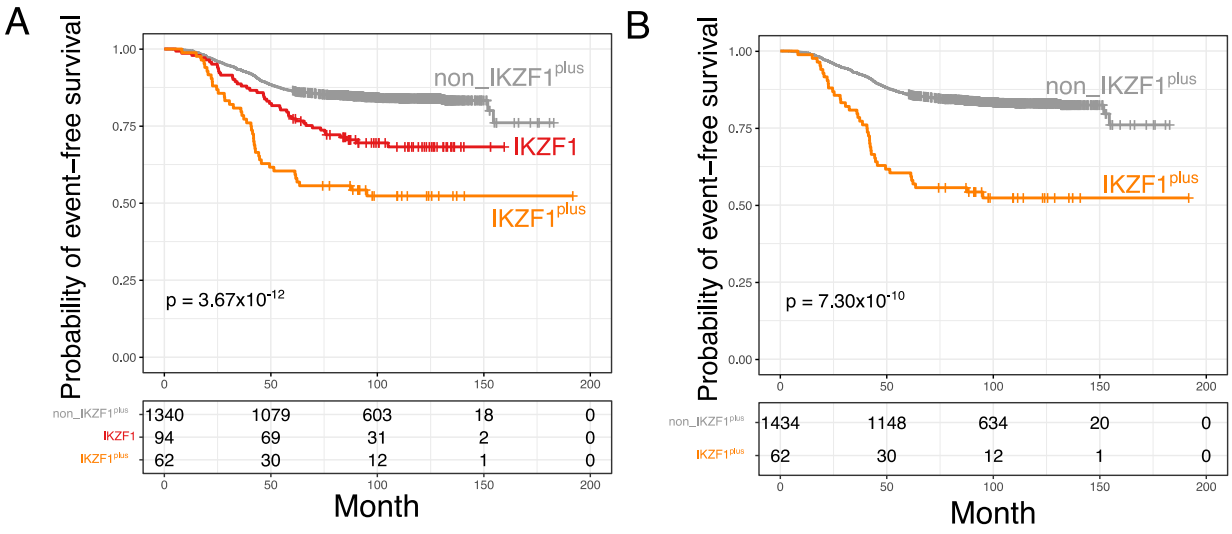
Supplementary Figure 15. *IKZF1* alterations and event free survival.

Weighted Kaplan-Meier (KM) curves of subclasses of *IKZF1* alteration in the entire MP2PRT study group. A. KM curves for groups classified based on types of exon deletions and SNV/indels. The samples with exon 4-5, exon 4-7 deletions or with missense mutations in the DNA binding zinc finger (ZnF) domains were classified as DEL_SNV_dominant_negative; the samples with the other types of exon deletions were classified as DEL_other. NoDel are samples without *IKZF1* deletions. The weighted cox-regression test results are shown on the right. B. KM curves for groups classified based on SNV/indel in the DNA binding ZnF domains. The samples with missense/nonsense/frameshift mutations in the DNA binding ZnF domains were classified as SNV_dominant_negative. NoDel are samples without *IKZF1* deletions.



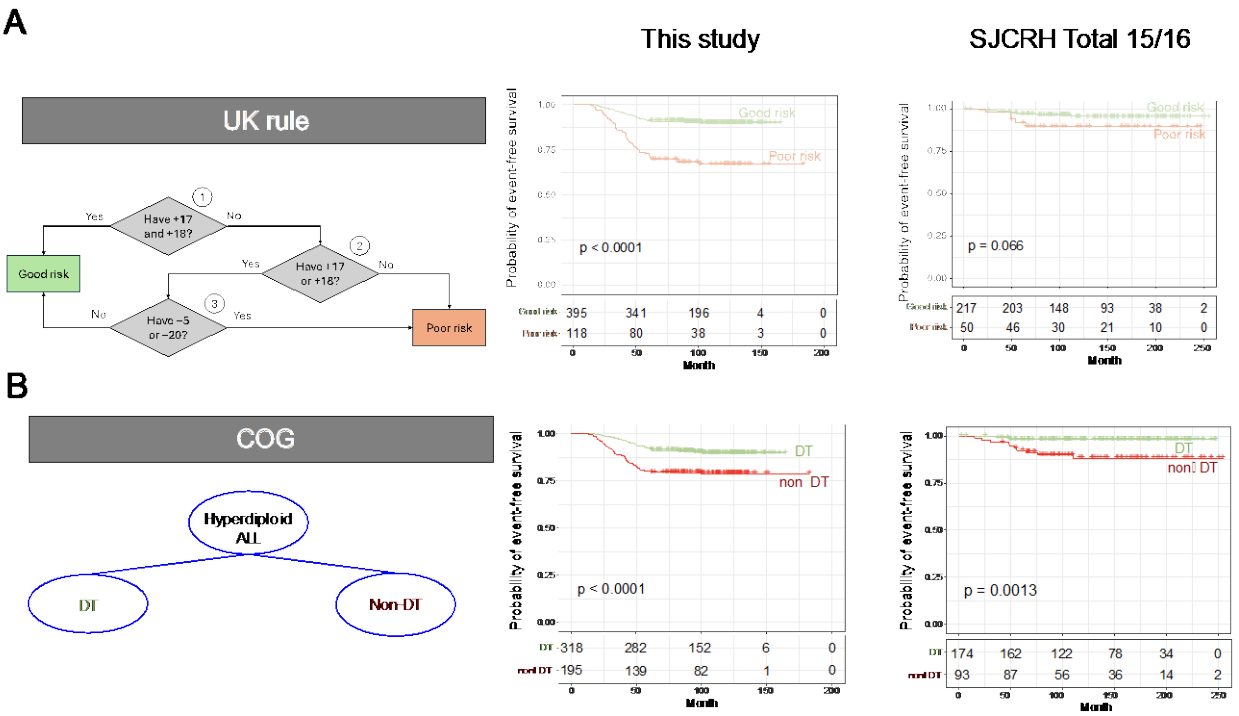
Supplementary Figure 16. Distribution of co-occurring deletions/rearrangements of IKZF1^{plus}.

Co-occurring deletion/rearrangements of IKZF1^{plus} in MP2PRT. **A**. MLPA-equivalent IKZF1^{plus}; **B**. MLPA-equivalent IKZF1^{plus}. The additional group, IKZF1, represents cases with *IKZF1* deletions but not the other co-occurring alterations; **C**. Genomically faithful IKZF1^{plus}; **D**. Genomically faithful IKZF1^{plus}. The additional group, IKZF1, represents cases with only *IKZF1* lesions but not the other co-occurring alterations.



Supplementary Figure 17. Weighted Kaplan-Meier curves of IKZF1^{plus} in MP2PRT.

A The weighted Kaplan-Meier curve for genomically accurate IKZF1^{plus} according to WGS data. **B**. The weighted Kaplan-Meier curve for a subclassification where the non-IKZF1^{plus} cases with *IKZF1* alterations were sub-classified as an independent group.



Supplementary Figure 18 Stratification of hyperdiploid ALL and outcome.

A, Left, UK stratification algorithm; middle, weighted KM curves for groups defined by the UK rules in the current study; right, KM curve for groups defined by the UK rules in SJ Total 15/16. **B**, Left, COG stratification algorithm; middle, weighted KM curves for groups defined by the COG rules in the current study; right, weighted KM curve for groups defined by the COG rules in SJ Total 15/16.

SUPPLEMENTARY REFERENCES

1. Mattano LA, Jr., Devidas M, Maloney KW, et al: Favorable Trisomies and ETV6-RUNX1 Predict Cure in Low-Risk B-Cell Acute Lymphoblastic Leukemia: Results From Children's Oncology Group Trial AALL0331. J Clin Oncol 39:1540-1552, 2021
2. Angiolillo AL, Schore RJ, Kairalla JA, et al: Excellent Outcomes With Reduced Frequency of Vincristine and Dexamethasone Pulses in Standard-Risk B-Lymphoblastic Leukemia: Results From Children's Oncology Group AALL0932. J Clin Oncol 39:1437-1447, 2021
3. Schore RJ, Angiolillo AL, Kairalla JA, et al: Outstanding outcomes with two low intensity regimens in children with low-risk B-ALL: a report from COG AALL0932. Leukemia 37:1375-1378, 2023
4. Larsen EC, Devidas M, Chen S, et al: Dexamethasone and High-Dose Methotrexate Improve Outcome for Children and Young Adults With High-Risk B-Acute Lymphoblastic Leukemia: A Report From Children's Oncology Group Study AALL0232. J Clin Oncol 34:2380-8, 2016
5. Burke MJ, Devidas M, Chen Z, et al: Outcomes in adolescent and young adult patients (16 to 30 years) compared to younger patients treated for high-risk B-lymphoblastic leukemia: report from Children's Oncology Group Study AALL0232. Leukemia 36:648-655, 2022
6. Salzer WL, Burke MJ, Devidas M, et al: Impact of Intrathecal Triple Therapy Versus Intrathecal Methotrexate on Disease-Free Survival for High-Risk B-Lymphoblastic Leukemia: Children's Oncology Group Study AALL1131. J Clin Oncol 38:2628-2638, 2020
7. Jeha S, Pei D, Choi J, et al: Improved CNS Control of Childhood Acute Lymphoblastic Leukemia Without Cranial Irradiation: St Jude Total Therapy Study 16. J Clin Oncol 37:3377-3391, 2019
8. Pui CH, Campana D, Pei D, et al: Treating childhood acute lymphoblastic leukemia without cranial irradiation. N Engl J Med 360:2730-41, 2009
9. Støer NC, Salim A, Bokenberger K, et al: Is the matched extreme case-control design more powerful than the nested case-control design? Stat Methods Med Res 28:1911-1923, 2019
10. Li H: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094-3100, 2018
11. Chen X, Gupta P, Wang J, et al: CONSERTING: integrating copy-number analysis with structural-variation detection. Nat Methods 12:527-30, 2015
12. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-60, 2009

13. Okonechnikov K, Conesa A, Garcia-Alcalde F: Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32:292-4, 2016
14. Cibulskis K, Lawrence MS, Carter SL, et al: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31:213-9, 2013
15. Larson DE, Harris CC, Chen K, et al: SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 28:311-7, 2012
16. Koboldt DC, Zhang Q, Larson DE, et al: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568-76, 2012
17. Fan Y, Xi L, Hughes DS, et al: MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol 17:178, 2016
18. Kim S, Scheffler K, Halpern AL, et al: Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods 15:591-594, 2018
19. Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164, 2010
20. Rausch T, Zichner T, Schlattl A, et al: DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333-i339, 2012
21. Chen X, Schulz-Trieglaff O, Shaw R, et al: Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32:1220-2, 2016
22. Cameron DL, Schroder J, Penington JS, et al: GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. Genome Res 27:2050-2060, 2017
23. Layer RM, Chiang C, Quinlan AR, Hall IM: LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15:R84, 2014
24. Chong Z, Ruan J, Gao M, et al: novoBreak: local assembly for breakpoint detection in cancer genomes. Nat Methods 14:65-67, 2017
25. Jeffares DC, Jolly C, Hoti M, et al: Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun 8:14061, 2017
26. Chiang C, Layer RM, Faust GG, et al: SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods 12:966-8, 2015
27. Blokzijl F, Janssen R, van Boxtel R, Cuppen E: MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med 10:33, 2018
28. Dobin A, Davis CA, Schlesinger F, et al: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15-21, 2013

29. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-40, 2010
30. Law CW, Chen Y, Shi W, Smyth GK: voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol 15:R29, 2014
31. Ritchie ME, Phipson B, Wu D, et al: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47, 2015
32. Jeha S, Choi J, Roberts KG, et al: Clinical significance of novel subtypes of acute lymphoblastic leukemia in the context of minimal residual disease-directed therapy. Blood Cancer Discov 2:326-337, 2021
33. Brady SW, Roberts KG, Gu Z, et al: The genomic landscape of pediatric acute lymphoblastic leukemia. Nat Genet 54:1376-1389, 2022
34. Tibshirani R, Hastie T, Narasimhan B, Chu G: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A 99:6567-72, 2002
35. Gu Z, Churchman ML, Roberts KG, et al: PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. Nat Genet 51:296-307, 2019
36. Stanulla M, Dagdan E, Zaliova M, et al: IKZF1(plus) Defines a New Minimal Residual Disease-Dependent Very-Poor Prognostic Profile in Pediatric B-Cell Precursor Acute Lymphoblastic Leukemia. J Clin Oncol 36:1240-1249, 2018
37. Therneau T, Atkinson B, Ripley B: Rpart: Recursive Partitioning. R Package Version 4.1-3., 2013
38. Kuhn M: Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28:1 - 26, 2008
39. Pounds S, Cheng C, Li S, et al: A genomic random interval model for statistical analysis of genomic lesion data. Bioinformatics 29:2088-95, 2013
40. Yekutieli D, Benjamini Y: Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference 82:171-196, 1999
41. Chen W, Wang S, Tithi SS, et al: A rare variant analysis framework using public genotype summary counts to prioritize disease-predisposition genes. Nat Commun 13:2592, 2022
42. Monsees GM, Tamimi RM, Kraft P: Genome-wide association scans for secondary traits using case-control samples. Genet Epidemiol 33:717-28, 2009
43. Maloney KW, Devidas M, Wang C, et al: Outcome in Children With Standard-Risk B-Cell Acute Lymphoblastic Leukemia: Results of Children's Oncology Group Trial AALL0331. J Clin Oncol 38:602-612, 2020

44. Secker-Walker LM, Berger R, Fenaux P, et al: Prognostic significance of the balanced t(1;19) and unbalanced der(19)t(1;19) translocations in acute lymphoblastic leukemia. Leukemia 6:363-9, 1992
45. Uckun FM, Sensel MG, Sather HN, et al: Clinical significance of translocation t(1;19) in childhood acute lymphoblastic leukemia in the context of contemporary therapies: a report from the Children's Cancer Group. J Clin Oncol 16:527-35, 1998
46. Andersen MK, Autio K, Barbany G, et al: Paediatric B-cell precursor acute lymphoblastic leukaemia with t(1;19)(q23;p13): clinical and cytogenetic characteristics of 47 cases from the Nordic countries treated according to NOPHO protocols. Br J Haematol 155:235-43, 2011
47. Felice MS, Gallego MS, Alonso CN, et al: Prognostic impact of t(1;19)/ TCF3-PBX1 in childhood acute lymphoblastic leukemia in the context of Berlin-Frankfurt-Munster-based protocols. Leuk Lymphoma 52:1215-21, 2011
48. Asai D, Imamura T, Yamashita Y, et al: Outcome of TCF3-PBX1 positive pediatric acute lymphoblastic leukemia patients in Japan: a collaborative study of Japan Association of Childhood Leukemia Study (JACLS) and Children's Cancer and Leukemia Study Group (CCLSG). Cancer Med 3:623-31, 2014
49. Paulsson K, Horvat A, Fioretos T, et al: Formation of der(19)t(1;19)(q23;p13) in acute lymphoblastic leukemia. Genes Chromosomes Cancer 42:144-8, 2005