Targeted capture sequencing

Sample preparation and valiant calling

For targeted sequencing, 200 ng of genomic DNA was enriched by liquid phase hybridization using the SureSelect custom kit, according to the manufacturer's protocol optimized for automated sample processing. We also sequenced 44 blood samples from cases in complete remission for excluding sequencing errors and single nucleotide polymorphisms (SNPs). Sequencing reads were aligned to the human genome reference (GRCh37) using Burrows-Wheeler Aligner, version 0.7.10 with default parameter settings. PCR duplicates were eliminated using picard-tools version 1.39 (http://picard.sourceforge.net/). Mutation calling was performed through our established pipeline, as previously reported.[1] The codes for our pipeline: Genomon2 algorithm is available at (http://genomon-project.github.io/GenomonPages/).

Adopt variants fulfilling the following criteria:
(i) Mapping Quality score ≥ 40
(ii) Base Quality score ≥ 20
(iii) Number of single nucleotide variants (SNVs) on the same read < 5
(iv) Number of insertions and deletions on the same read < 2
(v) Number of total reads ≥ 10
(vi) Number of variant reads ≥ 5
(vii) Variant allele frequency ≥ 0.02

Curation of the oncogenic variants

The detected candidate variations were filtered by excluding those with (i) synonymous or ambiguous (annotated as unknown by ANNOVAR) SNVs; (ii) known variants listed in SNP databases; (iii) variants only present in unidirectional reads; (iv) variants found in 44 samples at complete remission, with allele frequencies >0.2% on average; (vi) variants with <5 supporting reads in tumor samples; (v) Empirical Bayesian Mutation Calling P ≥ 0.0001 and (vi) all missense SNVs with a VAF of 0.4–0.6, unless registered as somatic mutations in hematopoietic and lymphoid tissue in the Catalogue of Somatic Mutations in Cancer (COSMIC) v82 databases. Finally, mapping errors were removed by visual inspection with IGV. For validation, 44 of 46 (96%) variations were confirmed by Sanger sequencing. Additionally, in 44 cases who have

germline control samples, all 54 mutations detected in the tumor samples were confirmed as somatic.

Copy number analysis

Copy number alterations (CNAs) were evaluated on the basis of sequencing data using our in-house pipeline CNACS.[2] In particular, the bait-library for targeted-capture sequencing was designed to capture 1,484 SNP sites for the measurement of genomic copy numbers. The codes for CNACS are available at https://github.com/papaemmelab/toil_cnacs. For each gene, estimated copy number < 1.6 and > 2.4 were considered to represent deletion and gain, respectively.

Detection of *IKZF1*^{plus}

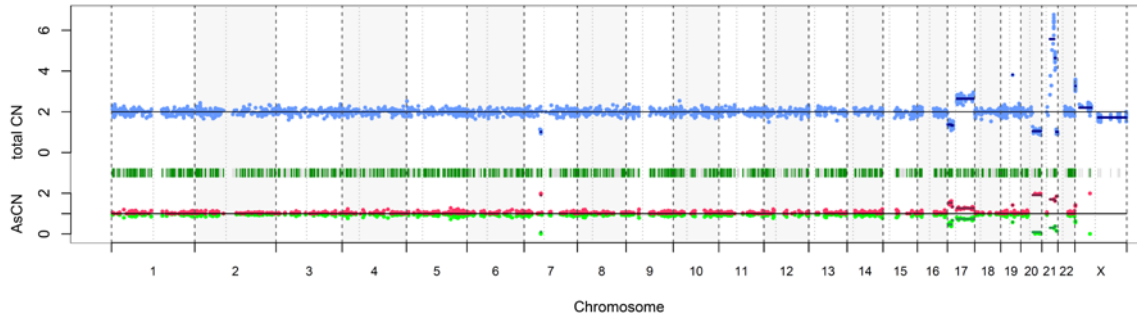
First, we defined *IKZF1*^{plus} group by targeted-sequencing based method, CNACS. Twenty-eight out of 84 cases with *IKZF1* deletion fulfilled *IKZF1*^{plus} criteria. In the 84 cases with *IKZF1* deletion, 67 cases (80%) were also analyzed by multiplex ligation-dependent probe amplification (MLPA) analyses (SALSA MLPA P335 ALL *IKZF1* kits; MRC-Holland, Amsterdam, the Netherlands) to assess copy number alterations of *PAX5*, *CDKN2A*, *CDKN2B* and Xp22.33/Yp11.31 region (whole chromosome X deletion was excluded in this study). Through this method, five patients were additionally defined as *IKZF1*^{plus} group.

Detection of structural variations

Detection of structural variations was performed by Genomon-SV (Y. Shiraishi, K.K., K.C., S. Miyano and S.O. et al., manuscript in preparation). Briefly, Genomon-SV uses information from chimeric reads (containing breakpoints) and discordant read pairs. For each candidate structural variation, it realigns reads to the assembled contig sequence containing the structural variation breakpoint (variant sequence). Putative structural variations were manually curated. For validation, 29 of 32 (90.6%) variations were confirmed by PCR and/or Sanger sequencing. The codes for Genomon-SV algorithm are available at https://github.com/Genomon-Project/GenomonSV.

Definition of intrachromosomal amplification of chromosome 21 (iAMP21)

iAMP21 was defined as that at least one signal of the total copy number of *RUNX1* gene exceeded five copies. We show a representative case with iAMP21.



Estimation of *TP53* mutated cell fraction

The estimated tumor cell fraction harboring *TP53* mutation (TCFmut) was calculated from total copy number (TCN) of the region, minor allele specific copy number (AsCN), and observed variant allele frequency (VAFobs) as follows;

TCFmut = VAFobs× (2-2× (1-AsCN)/(2-AsCN)) for deletions.

TCFmut = 1- AsCN for UPD

TCFmut = 2×VAFobs for regions without copy number changes

For gains, all mutations were identified in Hyperdiploid ALL. We tentatively assumed TCN as three and mutation occurred after trisomy formation. Tumor cell fractions can be estimated as

TCFmut = VAFobs×(2+2×(1-AsCN)/AsCN).

RNA analysis

Paired-end reads were aligned to the hg19 human genome assembly using STAR.[3] Fusion transcripts were detected by Genomon v.2.3.4 and filtered by excluding fusions (i) mapping to repetitive regions; (ii) with fewer than four spanning reads; (iii) that occurred out of frame; or (iv) had junctions not located at known exon–intron boundaries. The expression level for each RefSeq gene was calculated from mapped read counts using HTSeq and was normalized with the DESeq2 pipeline. To detect *ETV6-RUNX1*-like ALL in sequenced 116 cases, we performed hierarchical clustering using Ward's method for Euclidian distances with four *ETV6-RUNX1* fusion-positive ALL cases. Cluster stability was ascertained via consensus clustering with 1,000 iterations using the R package ConsensusClusterPlus. Six cases which were clustered with the four

ETV6-RUNX1 fusion-positive cases were defined as *ETV6-RUNX1*-like ALL. For analysis of Ph-like ALL, probe sets were selected from those (n = 263) previously used for clustering analysis to detect Ph-like ALL and they were reannotated to the updated RefSeq database, resulting in concatenation to a total of 187 RefSeq entries.[4,5] With these 187 probes, hierarchical clustering was performed on these 87 JACLS patients with three *BCR-ABL1* positive ALL cases. We defined patients as Ph-like ALL who were categorized with the three *BCR-ABL1* positive ALL cases. Ph-like ALL cases in TCCSG cohort were identified using expression array as previously reported.[6]

Refferences

1. Yoshizato T, Dumitriu B, Hosokawa K, et al. Somatic Mutations and Clonal Hematopoiesis in Aplastic Anemia. *N Engl J Med* 2015; **373**(1): 35-47.
2. Yoshizato T, Nannya Y, Atsuta Y, et al. Genetic abnormalities in myelodysplasia and secondary acute myeloid leukemia: impact on outcome of stem cell transplantation. *Blood* 2017; **129**(17): 2347-58.
3. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; **29**(1): 15-21.
4. Yasuda T, Tsuzuki S, Kawazu M, et al. Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. *Nat Genet* 2016; **48**(5): 569-74.
5. Harvey RC, Mullighan CG, Wang X, et al. Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome. *Blood* 2010; **116**(23): 4874-84.
6. Kobayashi K, Mitsui K, Ichikawa H, et al. ATF7IP as a novel PDGFRB fusion partner in acute lymphoblastic leukaemia in children. *Br J Haematol* 2014; **165**(6): 836-41.

Supplementary Figure Legends

Supplementary figure 1. Risk stratification of JACLS ALL02 protocol

The flow chart of patient classification in JACLS ALL02 protocol. For detailed information of treatment protocol in TCCSG, please refer to "Takahashi H, et al. Treatment outcome of children with acute lymphoblastic leukemia: the Tokyo Children's Cancer Study Group (TCCSG) Study L04-16. Int J Hematol. 2018. Abbreviations: PSL, prednisone; PPR, prednisone poor responder; PGR, prednisone good responder; WBC, white blood cell; SR, standard risk group; HR, high risk group; ER, extended high risk group; F, treatment failure group; pSR, provisional standard risk group; pHR, provisional high risk group; pER, provisional extended high risk group.

Supplementary figure 2. Genetic alterations in the total cohort.

Data are shown for 1,003 patients with pediatric B-precursor ALL who underwent targeted-capture sequencing. For details of captured genes and specific alterations, see Tables S3 and S4.

Supplementary figure 3. Representative deletions detected by targeted capture sequencing.

Frequent deletions in *ETV6*, *CDKN2A*, *PAX5*, *IKZF1*, and *ERG* are shown. Segmented copy number data are shown. Each row represents a patient; deleted regions are shown in blue.

Supplementary figure 4. Hyperdiploid ALL detected by targeted capture sequencing.

(A) Comparison of number of Hyperdiploid ALL cases detected by G-banding or targeted capture sequencing alone. (B) Box plots of white blood cell (WBC) count at diagnosis in patients with hyperdiploid ALL detected by sequence-based method alone or metaphase karyotyping method. In the box-and-whisker plots, the boxes indicates median and interquartle range, and whiskers denotes the range. The overlaid points are the observed WBC count for individual cases. (C) Kaplan–Meier curves for overall survival for patients with Hyperdiploid ALL detected by G-banding and those detected by targeted sequencing alone.

Supplementary figure 5. A representative copy number plot of hypodiploid ALL.

A representative copy number plot of hypodiploid ALL detected by sequencing based method.

Supplementary figure 6. Flow chart of classification of patients into 15 discrete subtypes.

Flow chart of classification of patients into 15 discrete subtypes.

Supplementary figure 7. Dendrograms of *ETV6-RUNX1*-like and Ph-like classification

Dendrograms of *ETV6-RUNX1*-like and Ph-like are shown. For details of clustering method, see supplementary appendix.

Supplementary figure 8. Kaplan-Meier curves for deletions and mutation in *TP53* in TARGET cohort.

Kaplan-Meier curves of event-free and overall survival for individuals in TARGET cohort with or without (A) *TP53* mutation and (B) *TP53* deletion.

Supplementary figure 9. Distribution of estimated fraction of *TP53* mutated cells.

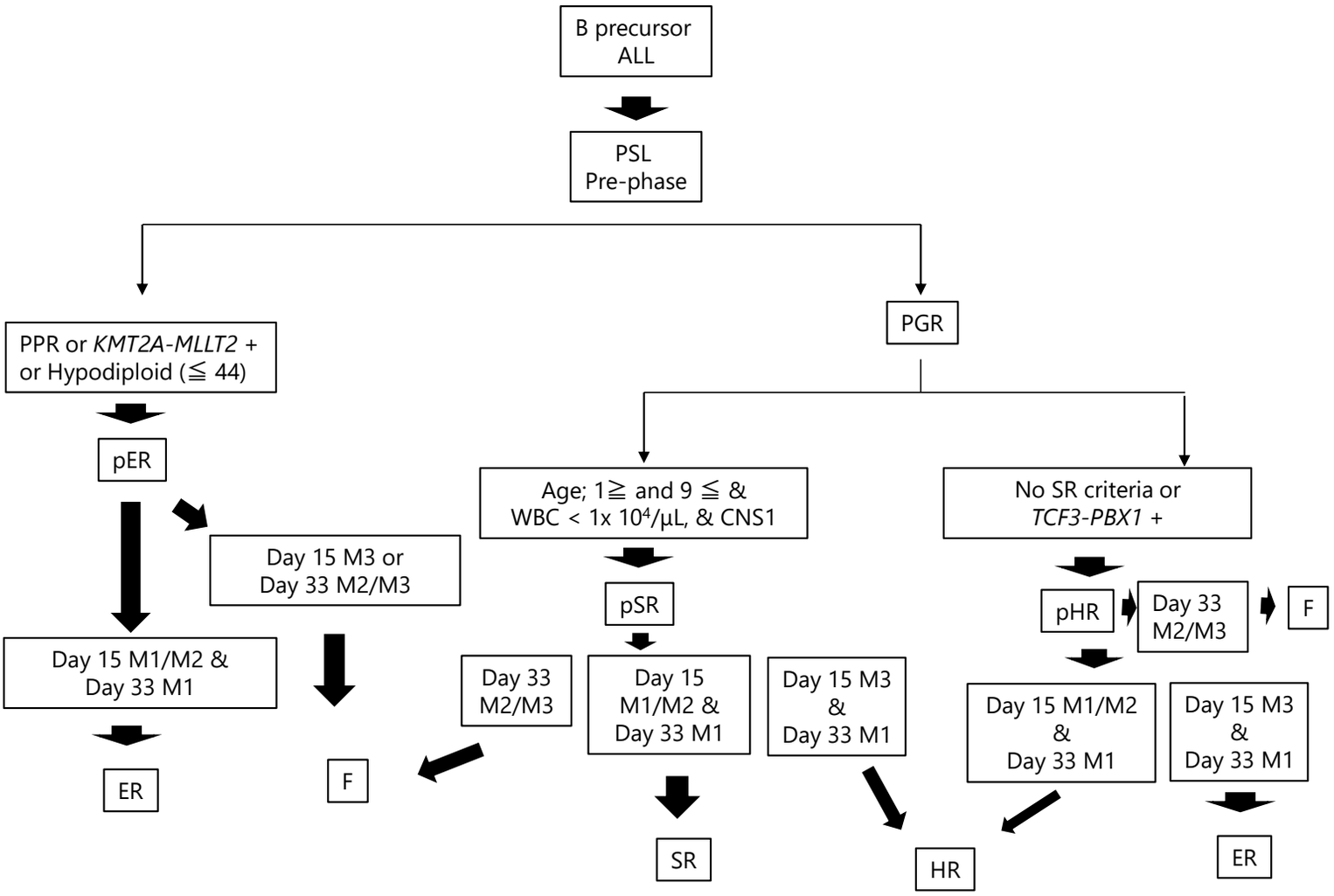
Histogram of distribution of the estimated cell fraction of *TP53* mutations. For estimtion of *TP53* mutated cell fraction, see supplementary appendix.

Supplementary figure 10. Correlation of clone size of *TP53* mutation with prognosis and NCI risk.

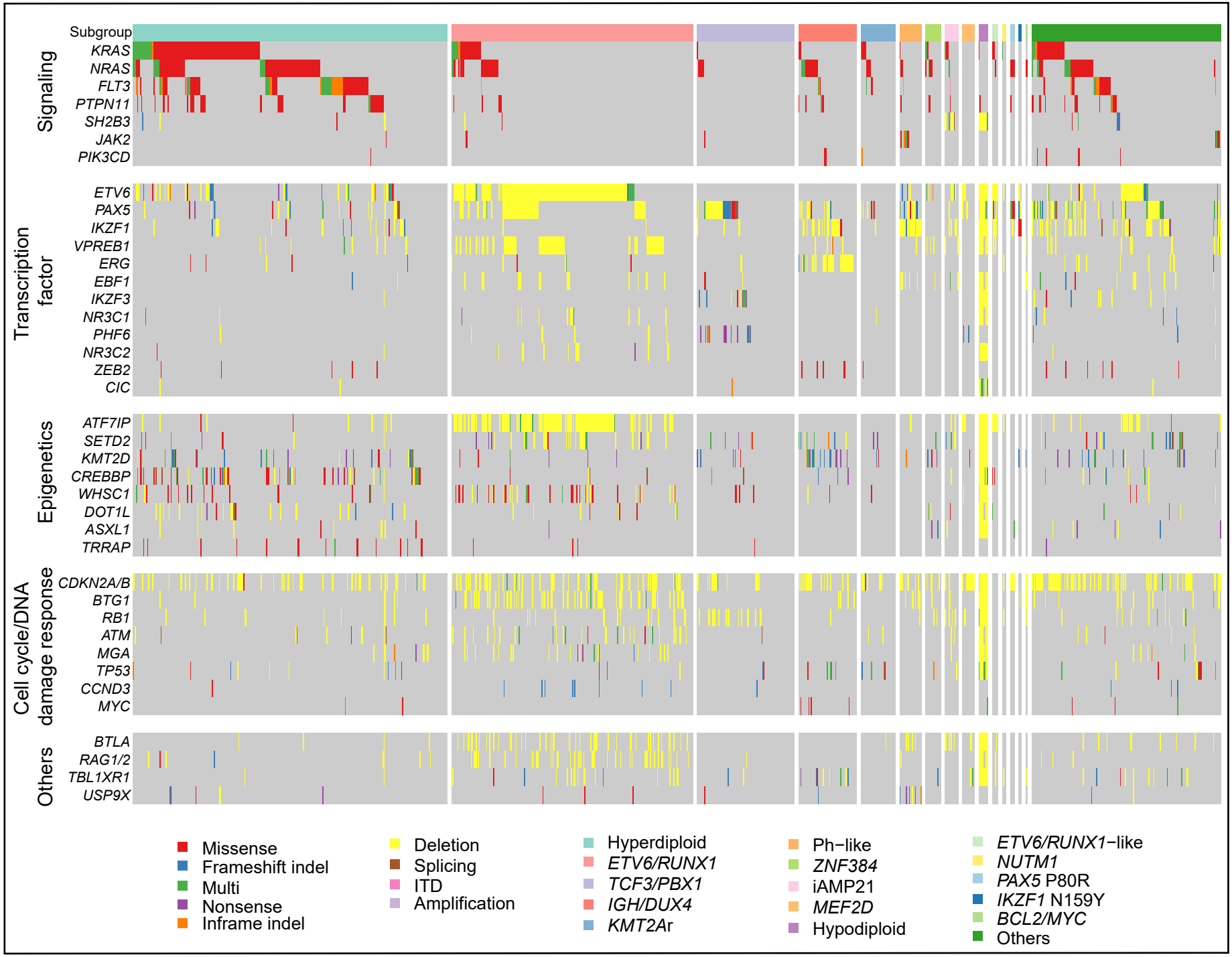
Box-and-whisker plots of estimated *TP53* mutated cell fraction according to the outcome of patients (A) and NCI risks (B). The boxes indicates median and

interquartle range, and whiskers denotes the range. For estimtion of *TP53* mutated cell fraction, see supplementary appendix.

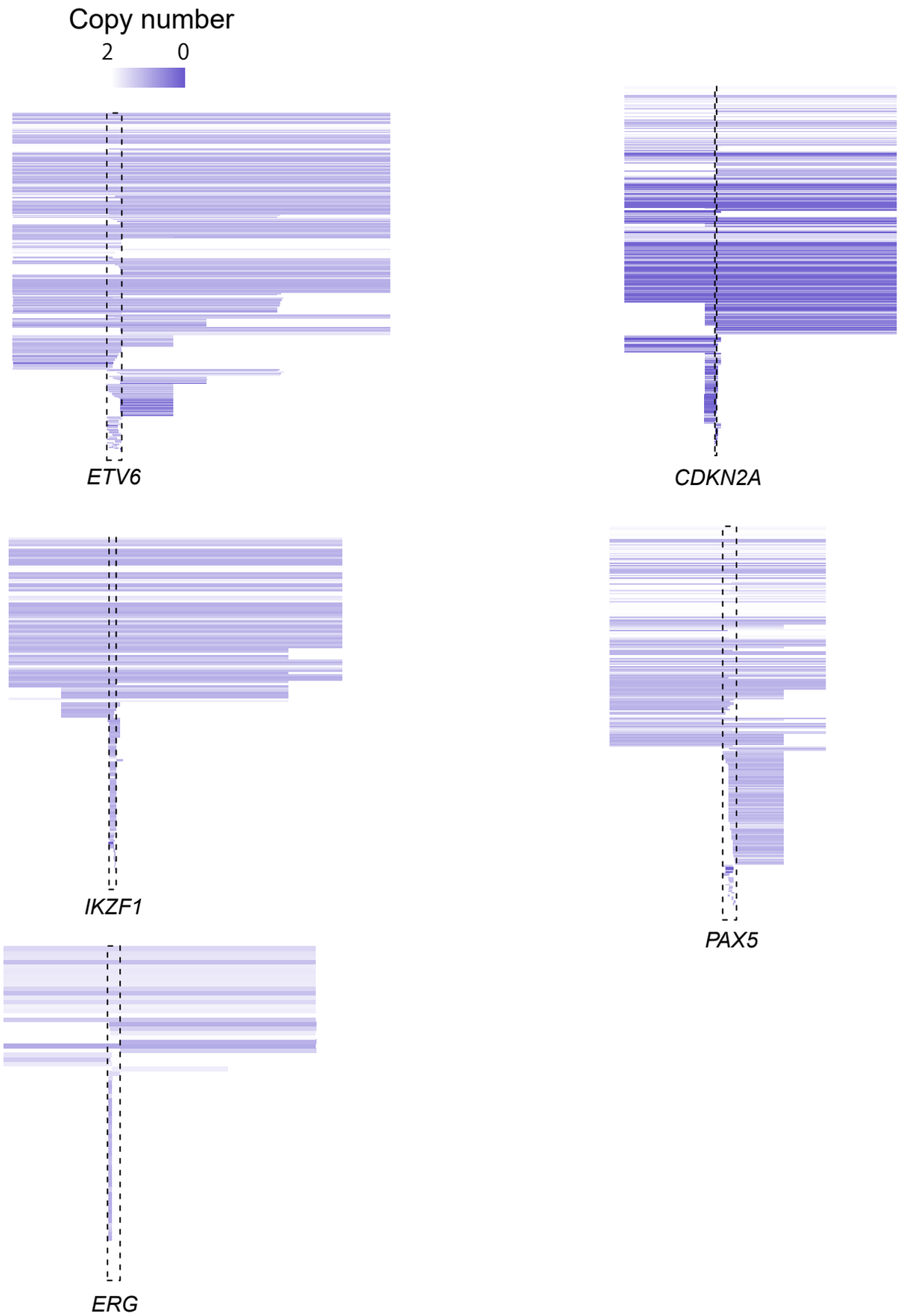
Supplementary figure 1



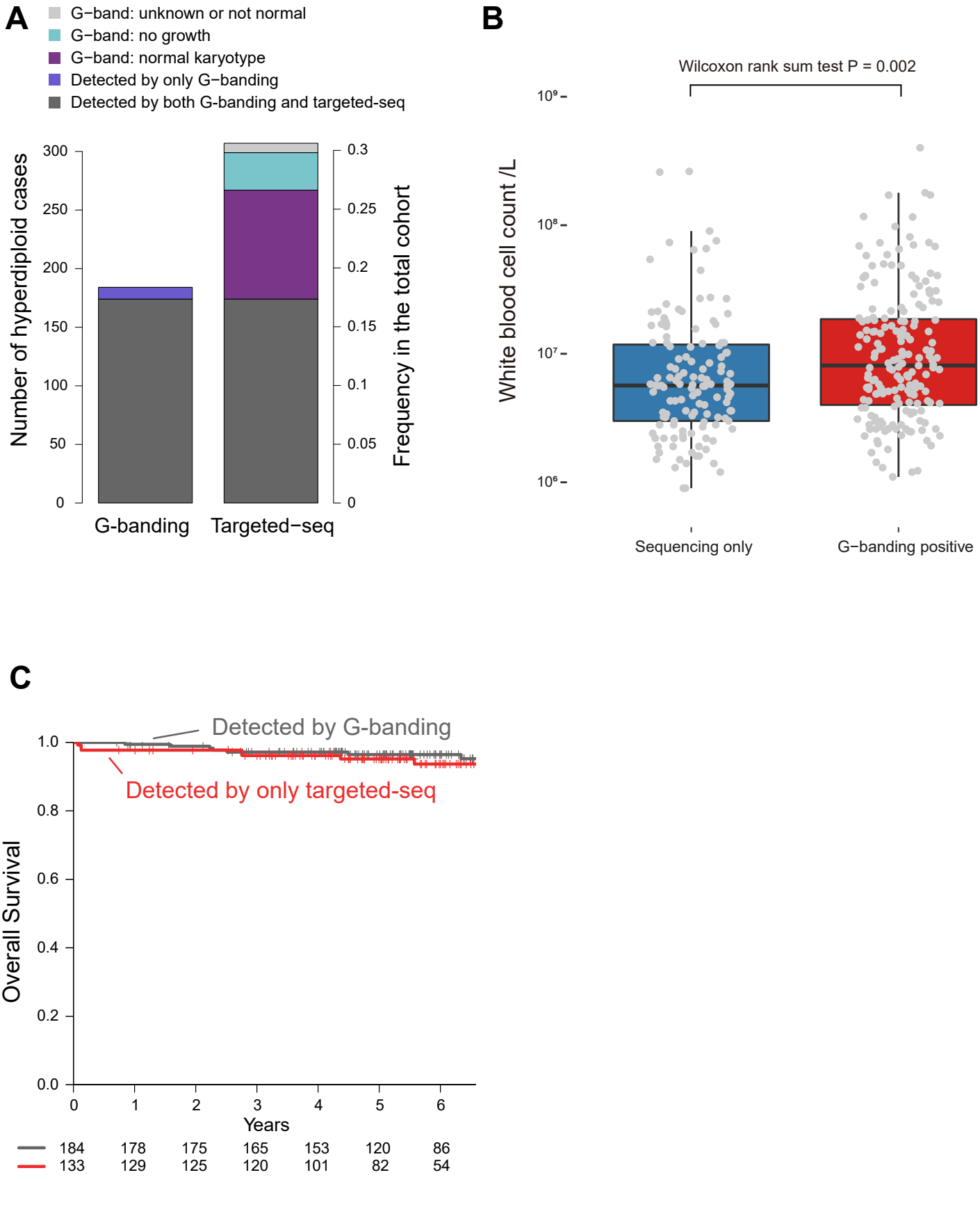
Supplementary figure 2



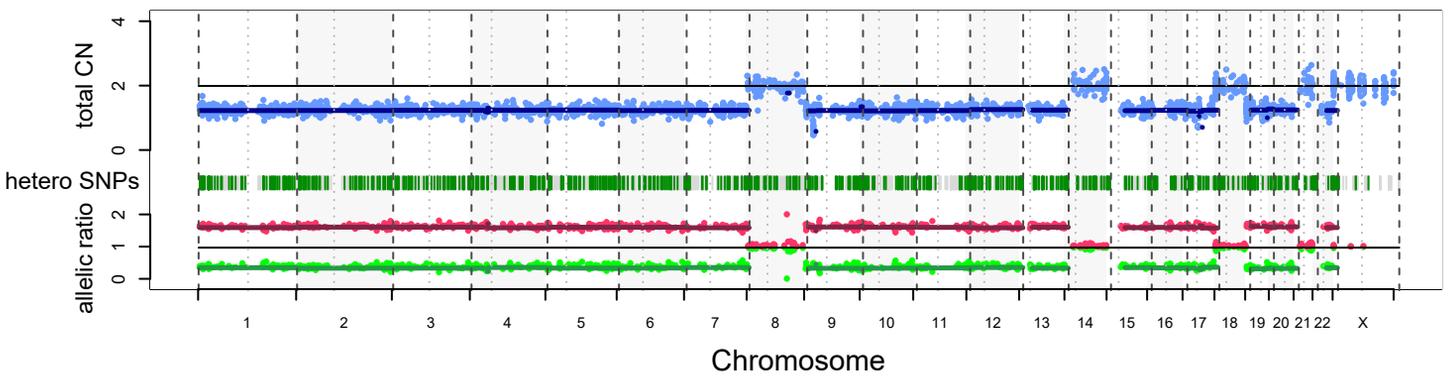
Supplementary figure 3



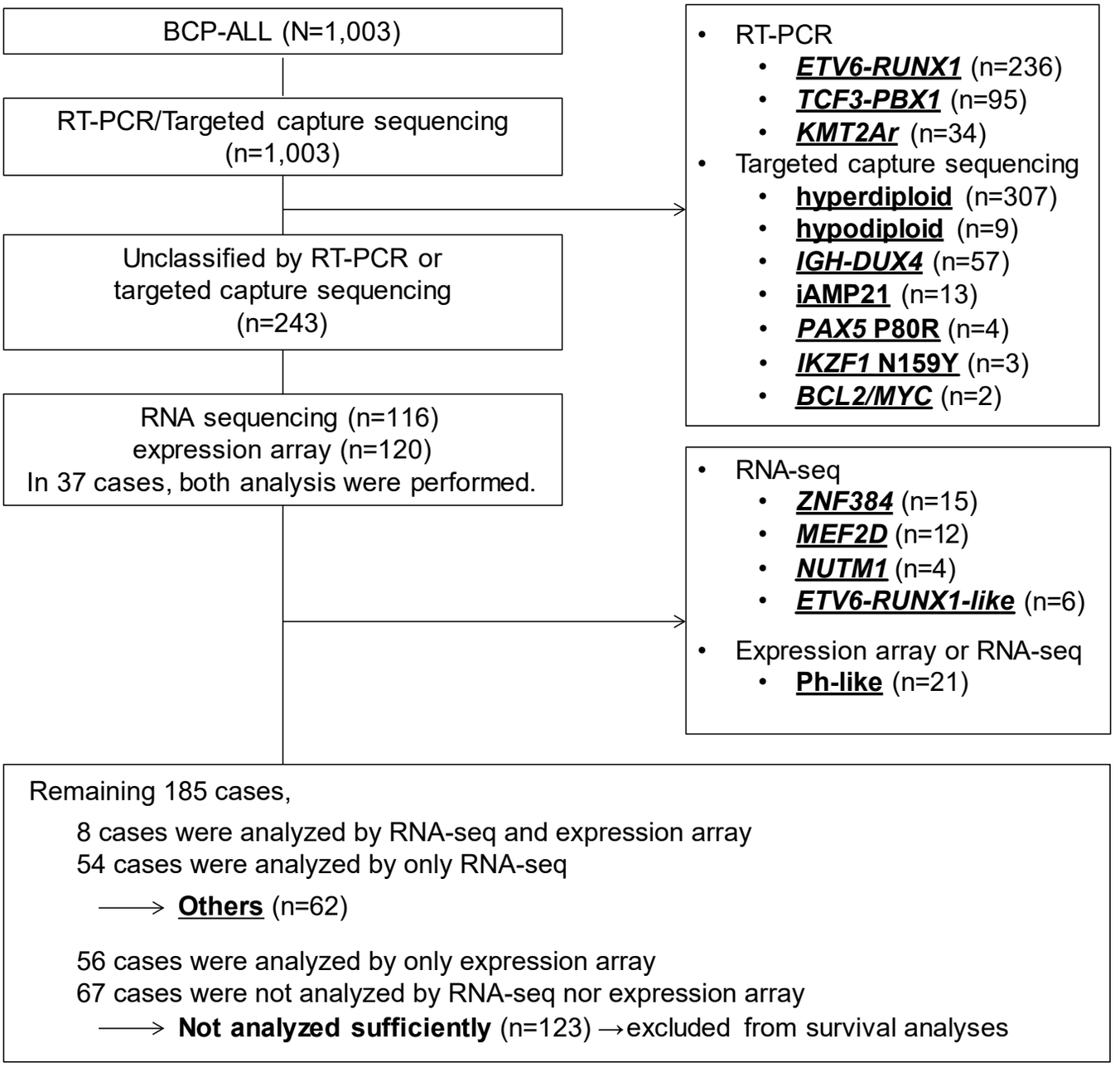
Supplementary figure 4



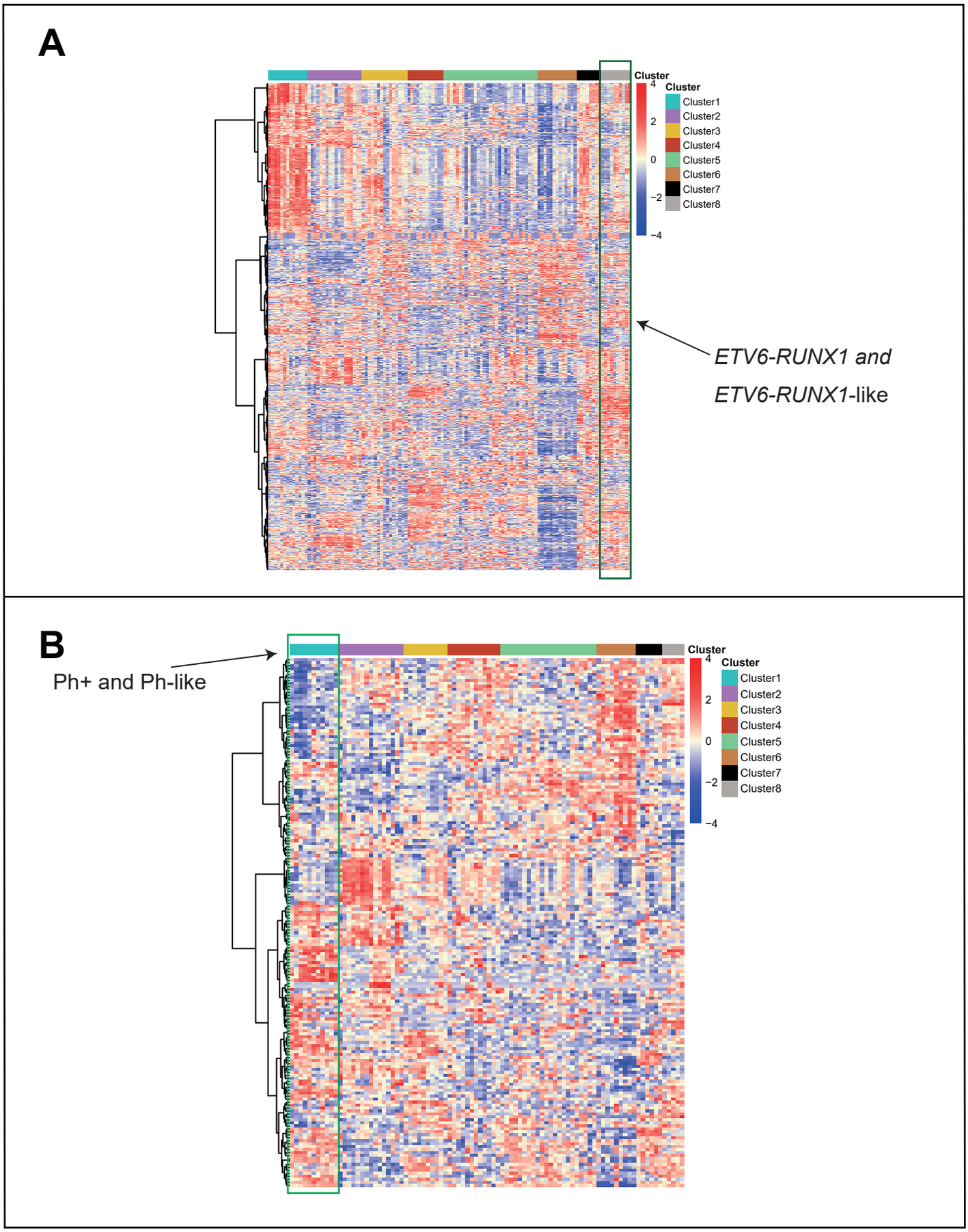
Supplementary figure 5



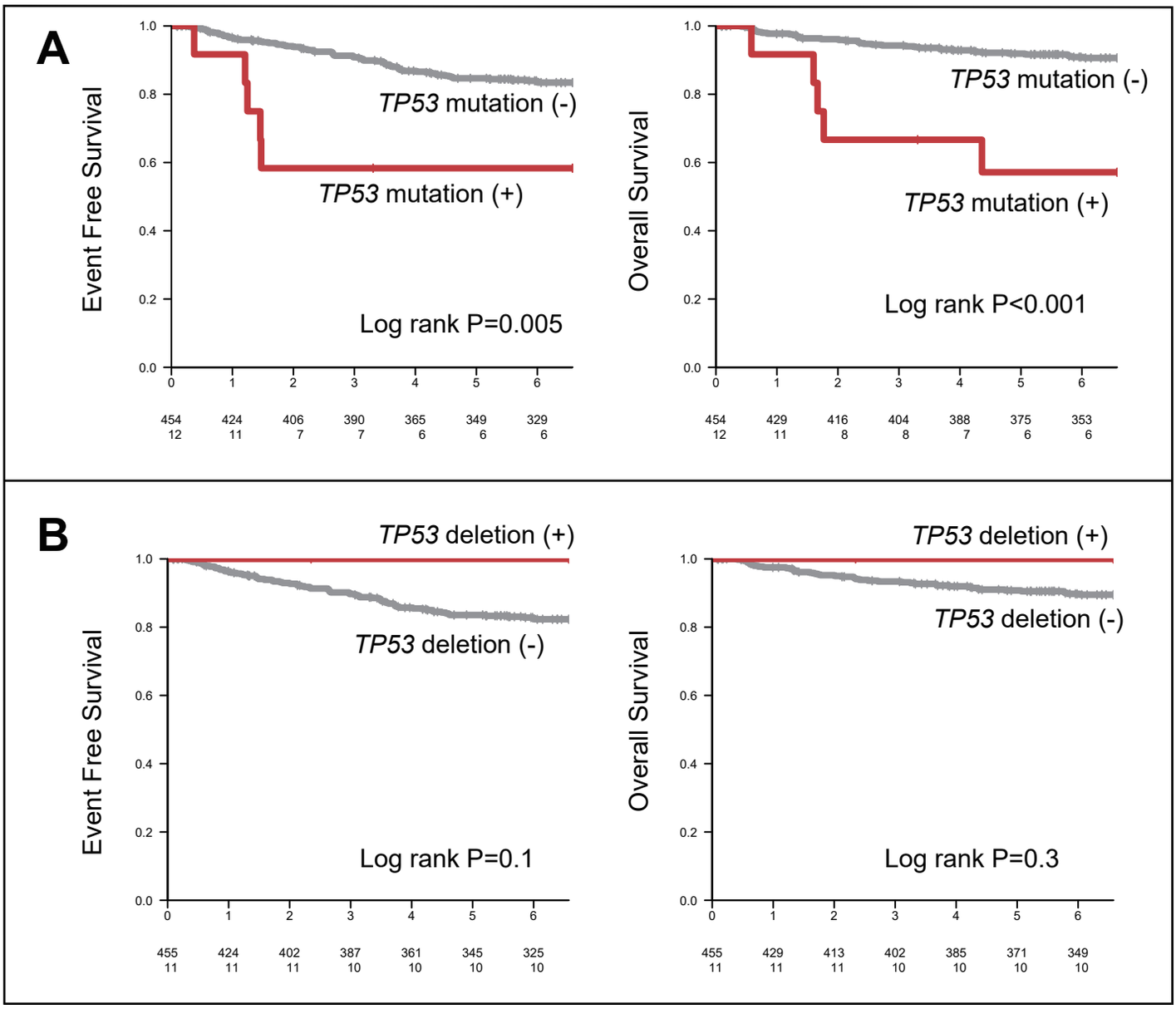
Supplementary figure 6



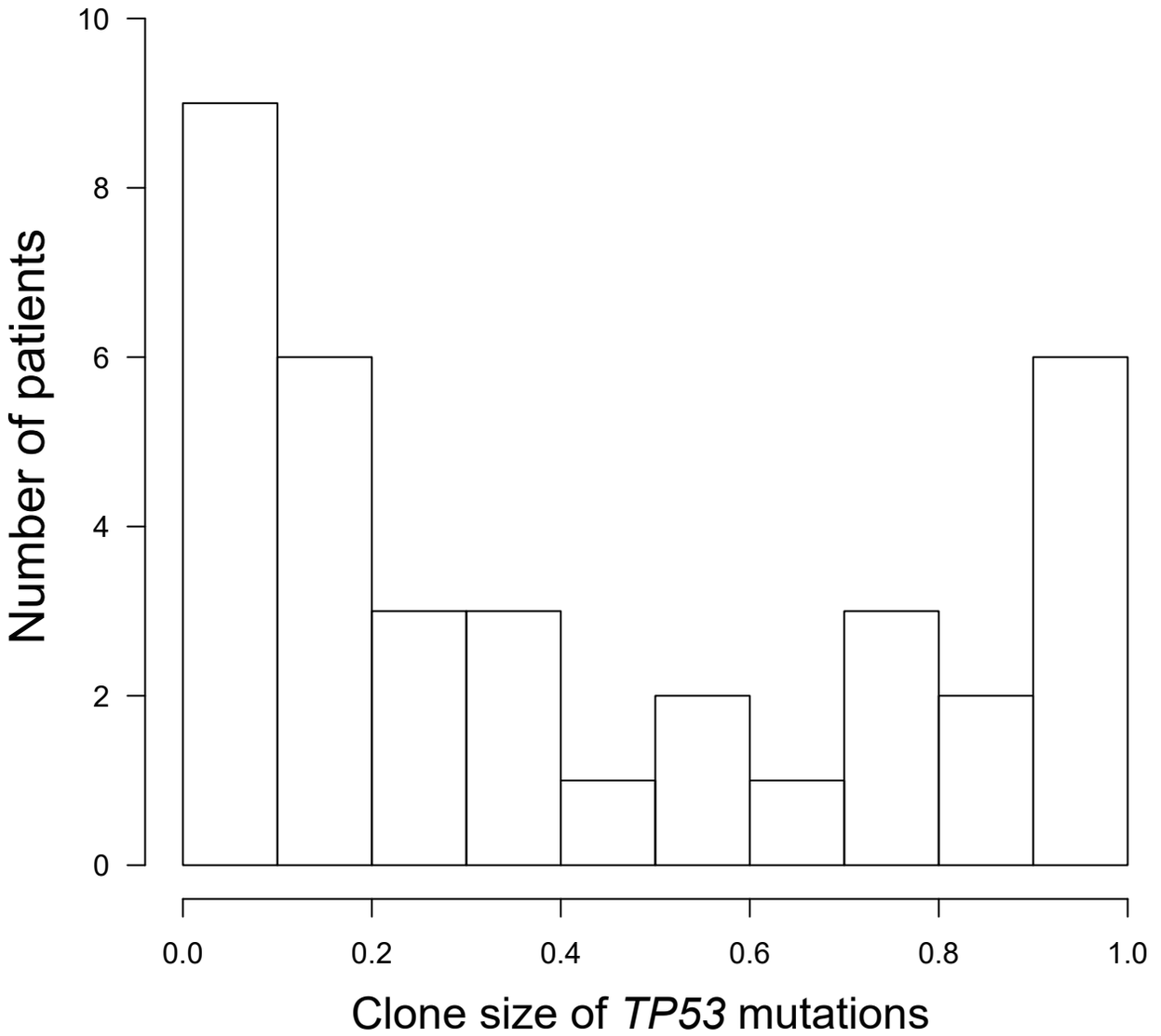
Supplementary figure 7



Supplementary figure 8

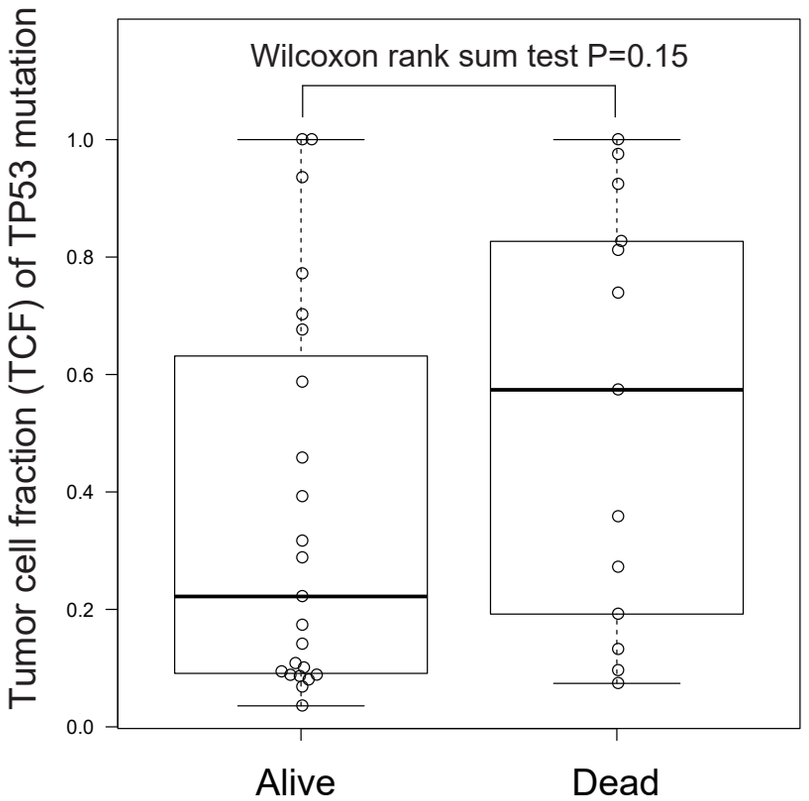


Supplementary figure 9



Supplementary figure 10

A



B