Supplemental Methods

Whole exome sequencing and variant calling on euploid and triploid chromosomes

Genetic data-preprocessing of whole exome sequencing data and germline variant calling were performed based on the Genome Analysis Tool Kit (GATK) best practices guidelines. [1] In summary, Burrows-Wheeler Aligner (BWA) 0.7 was used to align the FASTQ files to human reference genome 38 (hg38). [2] GATK 4.0 was then used to mark duplicates and base quality score recalibration in BAM files. Variant calling was performed with the HaplotypeCaller command in GVCF mode and joint variant calling across the entire cohort was subsequently performed using the GenotypeGVCFs command.

For non-chromosome 21 (*i.e.* euploid) chromosomes, we removed variants with a total read depth < 20, genotype quality (GQ) < 20, average GQ < 35, or missingness > 10%. [3] Variant Quality Score Recalibration (VQSR) was then applied with a truth sensitivity level of 99.5% for single nucleotide variants (SNVs) and 99.0% for insertion/deletion polymorphisms (indels), and variants that did not pass VQSR were removed. Given the high depth of sequencing, we used relatively stringent thresholds for further filtering of likely spurious variant calls, removing variants with quality by depth (QD) scores <5 and/or with an alternate allelic ratio <0.25.

For the triploid chromosome 21, following removal of variants with total read depth < 20, GQ < 20, average GQ < 35, or missingness >10%, we compared the variant calling accuracy of two different methods, GATK and Freebayes, [4] each of which were run in diploid or triploid ploidy mode. A higher number of chromosome 21 SNVs were called per sample by Freebayes than by GATK, in both diploid (Freebayes mean N=420 vs. GATK N=275) and triploid (Freebayes N=410 vs. GATK N=286) modes. We calculated the per sample ratio of transitions (Ti) to transversions (Tv) of SNVs called in each of the four modes to assess the quality of variant calls, [1] and found that variants called by GATK in diploid mode had the highest mean TiTv ratio (2.10), followed by GATK triploid mode (1.84), Freebayes triploid mode (1.66), and then Freebayes diploid mode (1.63) (**Figure S1**). This was likely due to Freebayes (in both modes) and GATK in triploid mode calling a higher number of variants with low alternate allele fraction, which were likely spurious, than GATK in diploid mode. Thus, we proceeded to generate variant calls using GATK Haplotype caller in diploid mode. Due to the relatively low number of variant calls on chromosome 21, we carried out hard filtering instead of VQSR for removal of likely spurious variants, with SNVs with QD < 5.0, FS > 60.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0 removed, and indels with QD < 5.0, FS > 200.0 and ReadPosRankSum < 20.0 removed.

For all chromosomes, variant annotation was performed using ANNOVAR, [5] which incorporated information for Gencode v26, [6] Genome Aggregation Database (gnomAD) 2.0.2 allele frequencies, [7] and ClinVar (07-01-2018). [8] BCFtools version 1.9 [9] was used to annotate variants with allele frequencies from the Trans-Omics for Precision Medicine (TOPMed) program Freeze 5 [10] and with Combined Annotation Dependent Depletion (CADD) Phred-scaled scores (version 1.4). [11]

Identification of pathogenic/likely pathogenic variants in predisposition genes

We limited our analyses to variants that are rare in the general population, by filtering out all SNVs or indels with allele frequencies >0.0001 (0.01%) in gnomAD or TOPMed. We retained only variants annotated as loss-of-function (stopgain, stoploss, frameshift insertion/deletion, splicing), as “Pathogenic” or “Likely pathogenic” (P/LP) in ClinVar, or that had a CADD Phred-scaled score ≥ 20. Variants annotated as “benign” or “likely benign” in ClinVar were removed. Further, we limited to variants overlapping genes in our candidate predisposition gene list (**Table S2**), which included 565 putative cancer-predisposition genes assessed in a previous sequencing study of germline variants in childhood cancer, [12] plus 9 genes at risk loci discovered in GWAS of ALL: *CEBPE*; *PIP4K2A*; *BMI1*; *IKZF3*, *GSDMB*, and *ORMDL3* (candidate genes at locus 17q12); *LHPP*; *ELK3*; and *SP4*; [13-19] in addition to *IKZF2*, which is somatically altered in some ALL tumors. [20] Additional ALL GWAS genes (*IKZF1*, *ARID5B*, *CDKN2A*, *GATA3*, *MYC*, *ERG*) [21-24] were already included in the 565 cancer-predisposition gene list.

Copy number analysis

To confirm trisomy 21 status in DS-ALL patients, we performed copy number analysis using two different tools, CNVkit [25] and CopywriteR, [26] both of which were designed for use with targeted or whole exome sequencing. Aligned BAM files generated by BWA were used as input for CNVkit, which uses both on-target and off-target reads to calculate $\log_2$ copy ratios across the genome, and for CopywriteR, which utilizes only the off-target reads for copy number analysis. In addition to trisomy 21 confirmation, we explored whether any copy number variants overlapped known ALL predisposition genes.

References

1. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics.* 2013;**43**:11.10.1-33.

2. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;**25**(14):1754-1760.

3. Carson AR, Smith EN, Matsui H, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics.* 2014;**15**:125-2105-15-125.

4. Garrison E, Marsh G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907.* 2012;.

5. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;**38**(16):e164.

6. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019;**47**(D1):D766-D773.

7. Karczewski KJ. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *Biorxiv.* **{Preprint]**.

8. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;**42**(Database issue):D980-5.

9. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;**27**(21):2987-2993.

10. Taliun D, Abecasis GR. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv.* 2019;.

11. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;**46**(3):310-315.

12. Zhang J, Walsh MF, Wu G, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med.* 2015;**373**(24):2336-2346.

13. Trevino LR, Yang W, French D, et al. Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet.* 2009;**41**(9):1001-1005.

14. Papaemmanuil E, Hosking FJ, Vijayakrishnan J, et al. Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet.* 2009;**41**(9):1006-1010.

15. Wiemels JL, Walsh KM, de Smith AJ, et al. GWAS in childhood acute lymphoblastic leukemia reveals novel genetic associations at chromosomes 17q12 and 8q24.21. *Nat Commun.* 2018;**9**(1):286-017-02596-9.

16. Vijayakrishnan J, Kumar R, Henrion MY, et al. A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukemia at 10q26.13 and 12q23.1. *Leukemia.* 2017;**31**(3):573-579.

17. Xu H, Yang W, Perez-Andreu V, et al. Novel susceptibility variants at 10p12.31-12.2 for childhood acute lymphoblastic leukemia in ethnically diverse populations. *J Natl Cancer Inst.* 2013;**105**(10):733-742.

18. Migliorini G, Fiege B, Hosking FJ, et al. Variation at 10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. *Blood.* 2013;**122**(19):3298-3307.

19. de Smith AJ, Walsh KM, Francis SS, et al. BMI1 enhancer polymorphism underlies chromosome 10p12.31 association with childhood acute lymphoblastic leukemia. *Int J Cancer.* 2018;**143**(11):2647-2658.

20. Holmfeldt L, Wei L, Diaz-Flores E, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. *Nat Genet.* 2013;**45**(3):242-252.

21. Sherborne AL, Hosking FJ, Prasad RB, et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nat Genet.* 2010;**42**(6):492-494.

22. Wiemels JL, Walsh KM, de Smith AJ, et al. GWAS in childhood acute lymphoblastic leukemia reveals novel genetic associations at chromosomes 17q12 and 8q24.21. *Nat Commun.* 2018;**9**(1):286-017-02596-9.

23. Qian M, Xu H, Perez-Andreu V, et al. Novel susceptibility variants at the ERG locus for childhood acute lymphoblastic leukemia in Hispanics. *Blood.* 2019;**133**(7):724-729.

24. de Smith AJ, Walsh KM, Morimoto LM, et al. Heritable variation at the chromosome 21 gene *ERG* is associated with acute lymphoblastic leukemia risk in children with and without Down syndrome. *Leukemia.* 2019;**33**(11):2746-2751.

25. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol.* 2016;**12**(4):e1004873.

26. Kuilman T, Velds A, Kemper K, et al. CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol.* 2015;**16**:49-015-0617-1.

Supplemental Figure and Table Legends

Figure S1 – TiTv ratios for chromosome 21 variant calls using different calling methods. Boxplots displaying TiTv ratios for variant calls on the trisomic chromosome 21 generated from whole exome sequencing data (BAM files) from the 73 DS-ALL patients using two different calling algorithms, GATK and Freebayes, each in either diploid or triploid ploidy mode. Median and mean TiTv ratios across the 73 patients are displayed by horizontal black lines and black crosses, respectively. Boxplots were generated using BoxPlotR (http://shiny.chemgrid.org/boxplotr/).

Figure S2 – Chromosome 21 copy number analysis revealing partial trisomy 21 in a DS-ALL patient. We performed copy number analysis using whole exome sequencing data to confirm trisomy 21 status in the 73 DS-ALL patients. Panel **A** shows $\log_2$ ratio values across chromosome 21 for several full trisomy 21 patients and one apparent partial trisomy 21 patient (23168), generated using the CNVkit software. Panel **B** shows the chromosome 21 $\log_2$ ratios for the partial trisomy 21 patient, generated using the CopywriteR tool.

Figure S3 – Pathogenic germline *IKZF1* variant in DS-ALL patient. Patient 21869 harbored a germline p.Arg162Trp variant in *IKZF1*, which was confirmed by PCR and Sanger sequencing (**A**). This variant is located at a hotspot codon for germline predisposition to B-ALL, as additional variants in this codon, p.Arg162Pro and p.Arg162Gln, have been described in individuals with childhood B-ALL (**B**). The

mutation "lollipop plot", generated using the St. Jude Cloud Protein Paint software (https://pecan.stjude.cloud/proteinpaint), displays all rare germline variants (N=22 amino acid changes) confirmed to be functional in B-ALL patients in: i) ~5,000 sporadic B-ALL patients from the St. Jude Children's Research Hospital/Children's Oncology Group study (Churchman *et al*. 2018), ii) from ClinVar (*i.e.* p.Arg162Gln), and iii) in our DS-ALL patient. Codon 162 is the only one reported with multiple different amino acid changes that predispose ALL.

Figure S4 – Sanger sequencing validation of rare germline variants in predisposition genes in DS-ALL patients. Screenshots of Sanger sequencing chromatograms showing confirmation of likely pathogenic (*IKZF1*, *NBN*, *AKAP9*, *RTEL1*, *FOXP1*) and possibly pathogenic (*ARID5B*, *APC*) germline variants DS-ALL patients.

Figure S5 – Integrative Genomics Viewer (IGV) screenshot of germline *MLLT1* variant. The *MLLT1* germline p.Arg473Gln variant was supported by 61 mutant reads out of 129 total reads, with approximately equal alternate calls in forward and reverse reads. This IGV screenshot displays 13 mutant reads out of a total of 32 reads.

Table S1 – Demographic and clinical data for 73 DS-ALL patients in the International Study of Down Syndrome Acute Leukemia (IS-DSAL) included in this whole exome sequencing study.

Table S2 – Candidate cancer predisposition gene list.

Table S3 – Rare and predicted functional germline variants in cancer-related genes and relevant variant annotations, in 73 DS-ALL patients in the IS-DSAL.

Figure S1

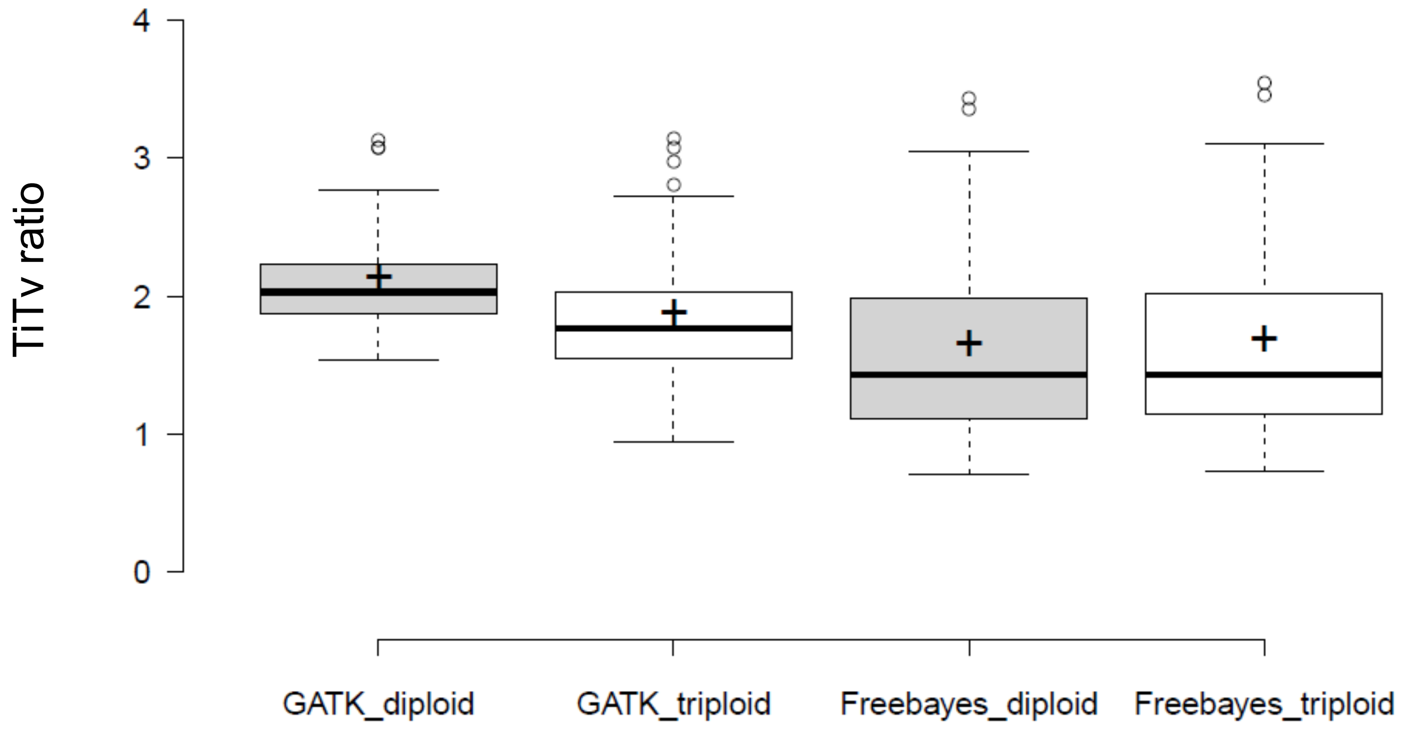

TiTv ratio
0
1
2
3
4
GATK_diploid
GATK_triploid
Freebayes_diploid
Freebayes_triploid

Figure S2

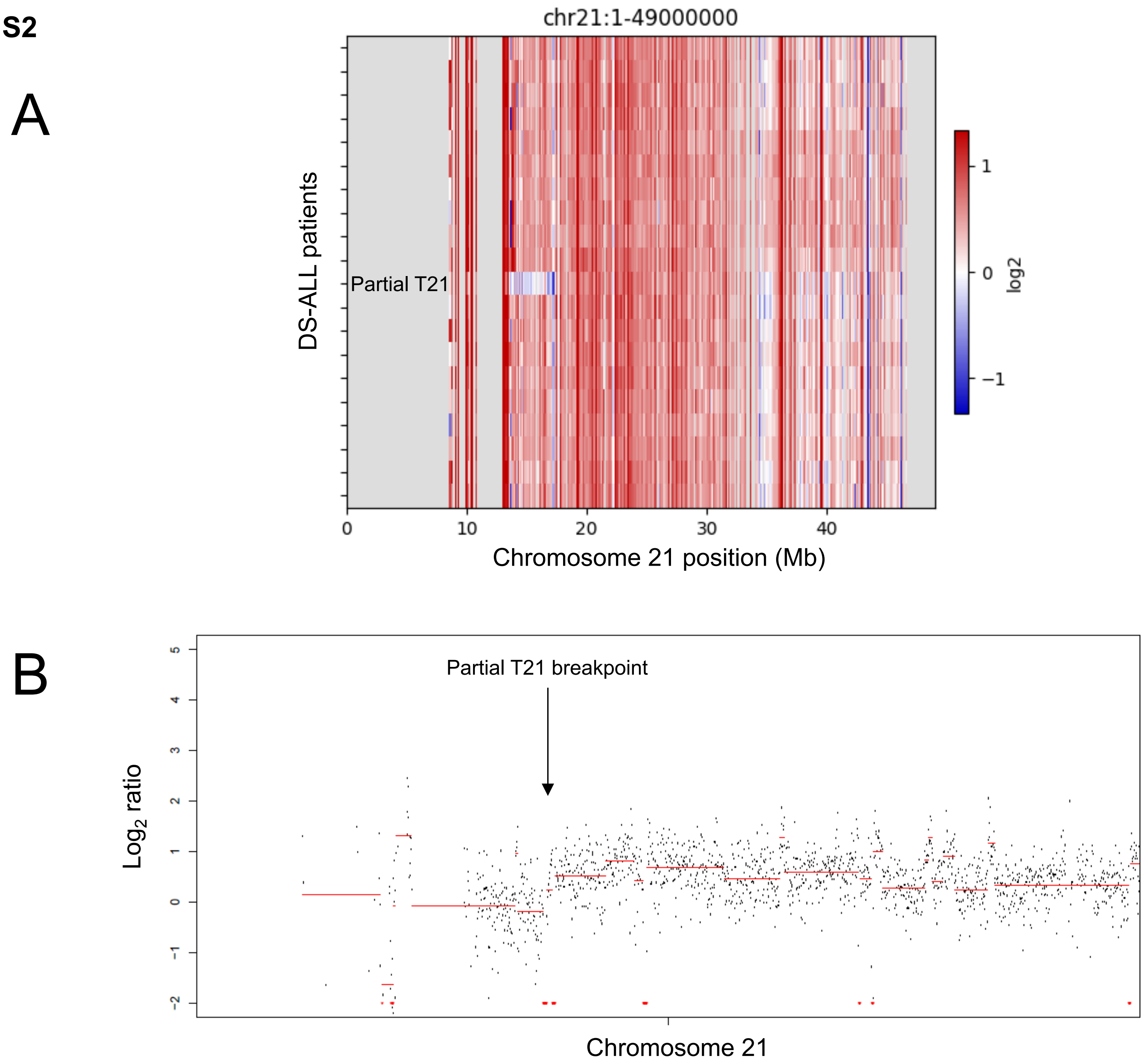
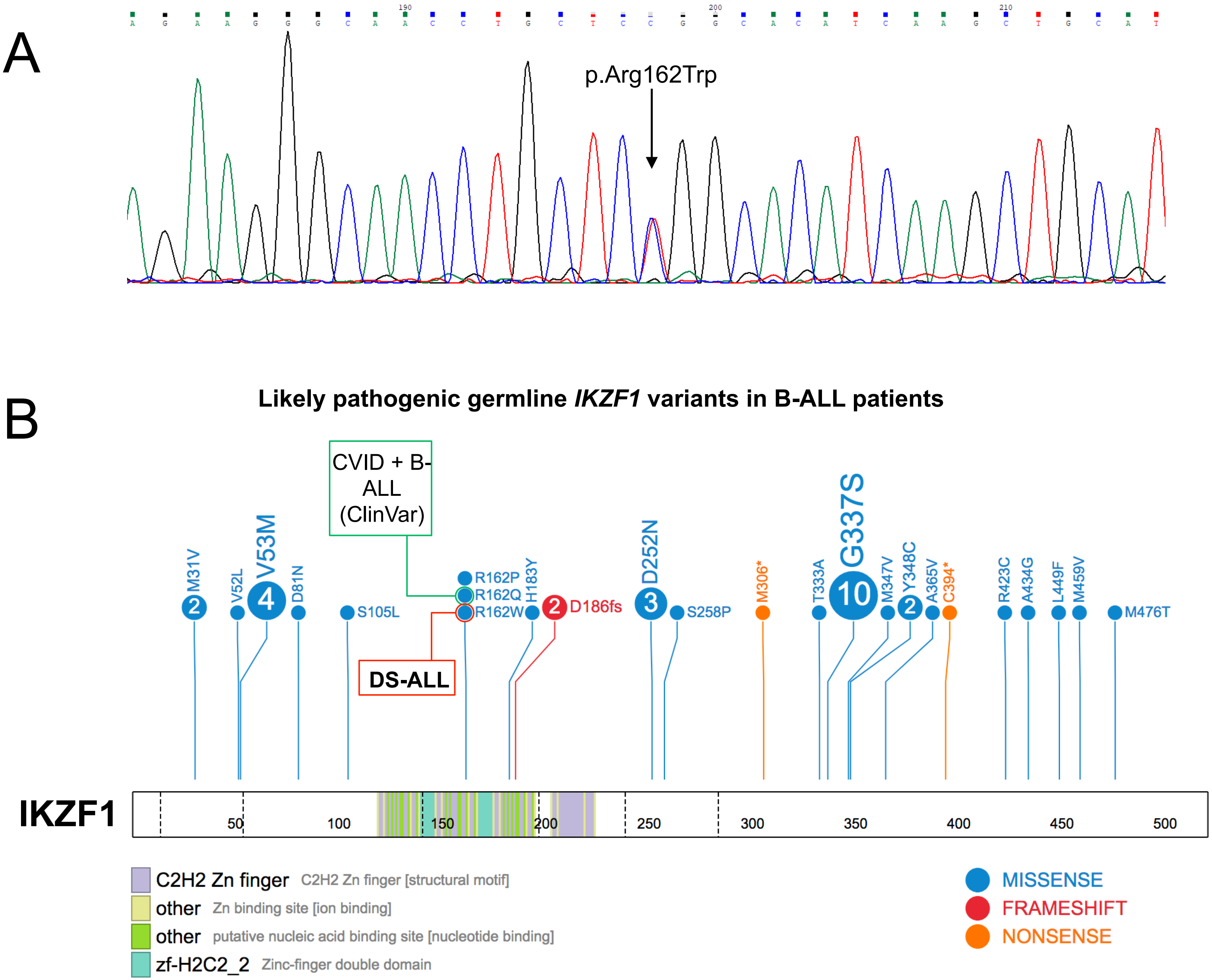
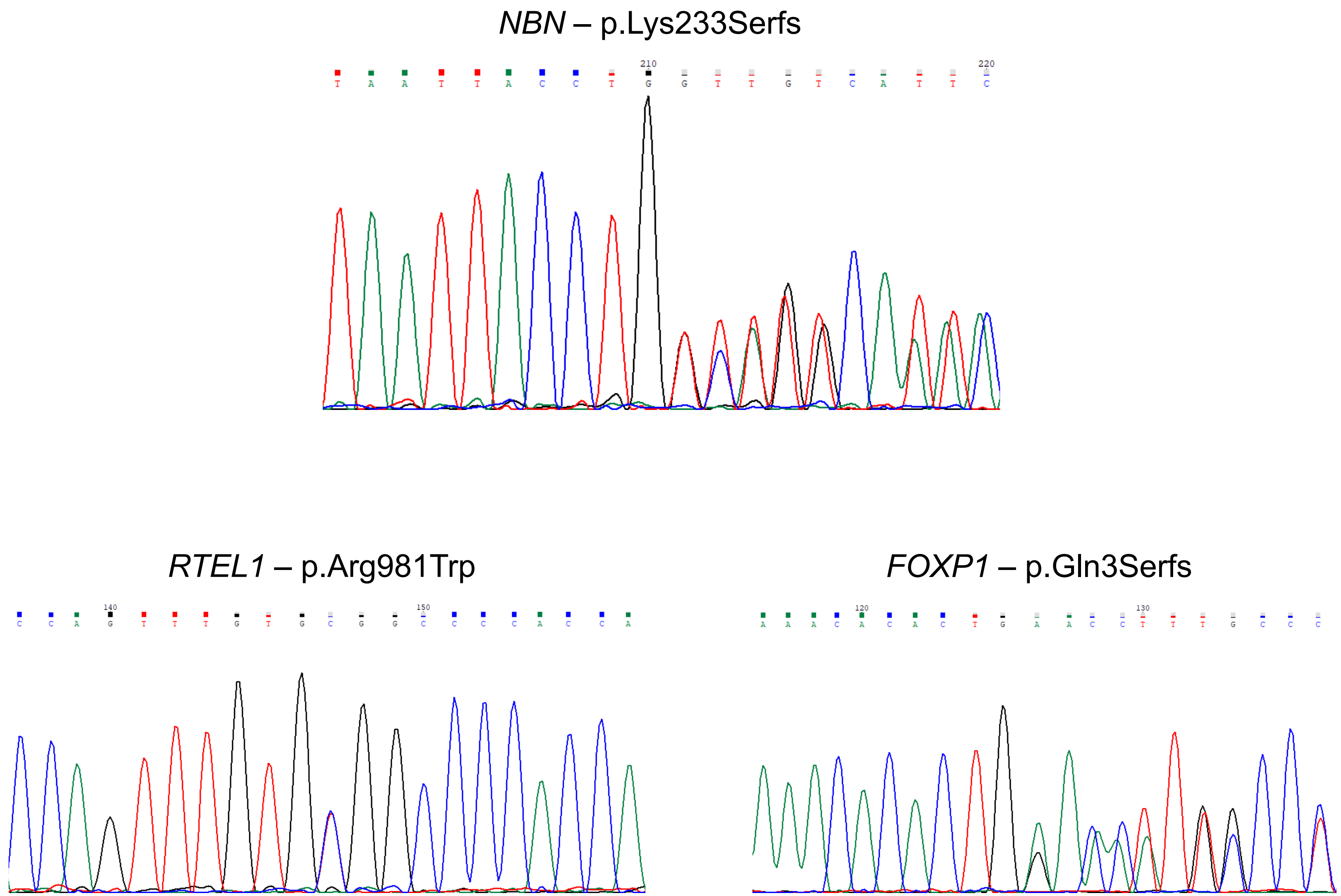

A
chr21:1-49000000
DS-ALL patients
Partial T21
0
10
20
30
40
Chromosome 21 position (Mb)
1
0
−1
log2
B
Partial T21 breakpoint
Log_2 ratio
5
4
3
2
1
0
-1
-2
Chromosome 21

Figure S3



Likely pathogenic germline *IKZF1* variants in B-ALL patients

Figure S4



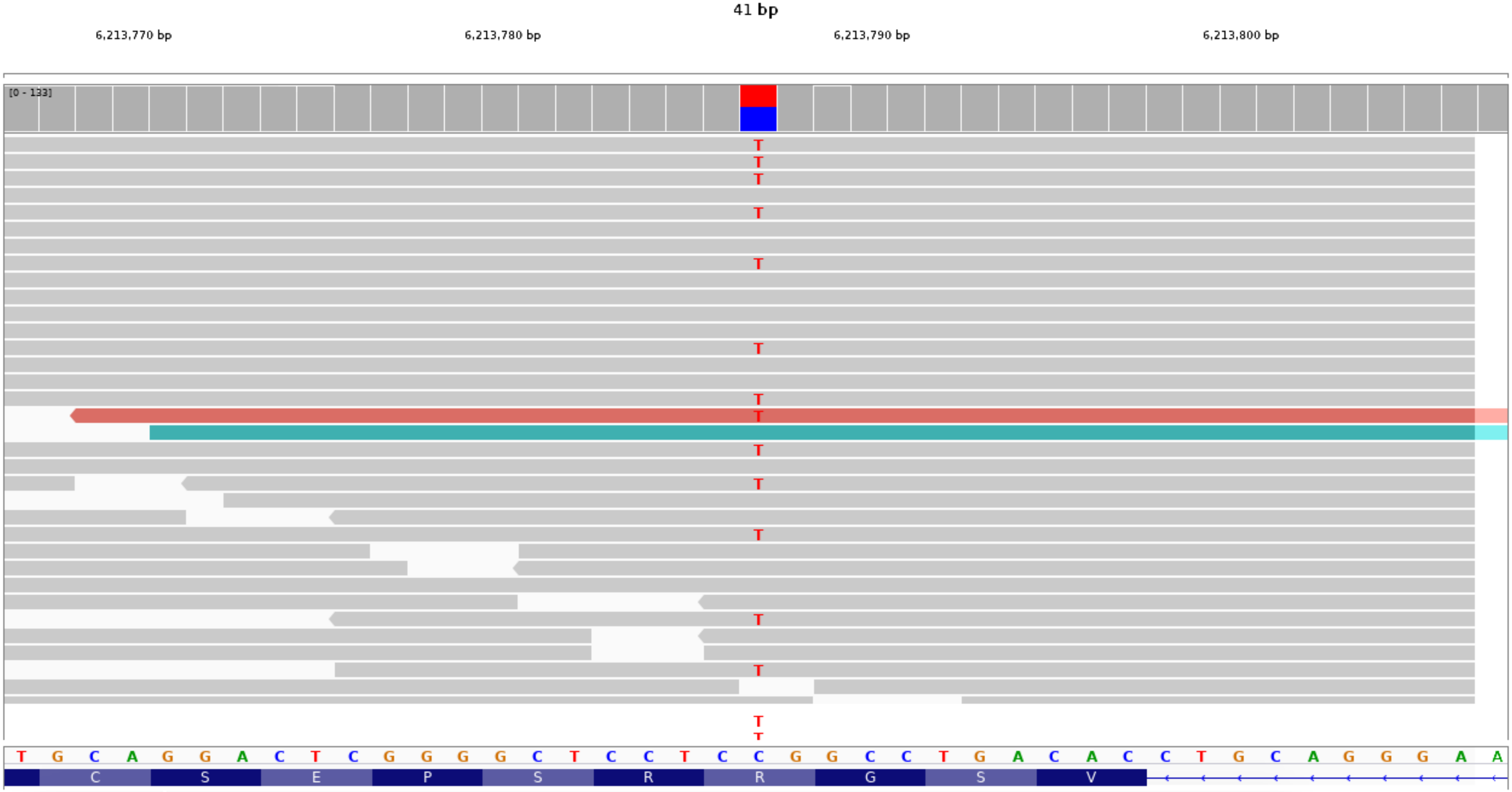
NBN – p.Lys233Serfs

RTEL1 – p.Arg981Trp

FOXP1 – p.Gln3Serfs

Figure S5

MLLT1 p.Arg473Gln variant


41 bp
6,213,770 bp
6,213,780 bp
6,213,790 bp
6,213,800 bp
[0 - 133]
T G C A G G A C T C G G G G C T C C T C C G G C C T G A C A C C T G C A G G G A A
C S E P S R R G S V