SUPPLEMENTARY DATA

SUPPLEMENTAL METHODS

Transcriptome sequencing (RNA-seq) and pathway analysis

Sequencing reads from total stranded protocol were mapped to the GRCh37 human genome reference by STAR[1] (version 2.5.3a) through the suggested two-pass mapping pipeline, with reads mapped to more than 20 loci considered as unmapped. Gene expression was quantified using RSEM[2] (v1.3.0) against Ensembl gene annotation (v75) (http://www.ensembl.org/). All the samples were sequenced with RefSeq coding region covered with 30-fold coverage ≥15% (median ± standard deviation, 38.3±6%). Sequencing reads from low quality or low input protocols were preprocessed by removing reads with low complexity (Prinseq v0.20.4)[3] or mapped to ribosome RNAs and trimmed with trimgalore (v0.4.4) (https://github.com/FelixKrueger/TrimGalore), followed by STAR mapping and RSEM gene quantification as applied to total stranded libraries. FusionCatcher[4] was used to detect fusions, and all the reported rearrangements were manually reviewed to keep the reliable ones. Due to the complexity of *IGH* and *DUX4* rearrangements, some of the fusions were manually rescued by checking the aligned reads within IGV browser.[5] In addition, expression of two highly repetitive genes, *DUX4* and *IGH* was also quantified with kallisto (v43)[6] using all transcript sequences regardless of gene location in the genome. DUX4 rearrangement was confirmed with high DUX4 expression. Differential gene expression analysis was carried out using Limma Bioconductor R package with voom transformation of read counts from RSEM.[7] Differentially expressed genes were selected using p-value of 0.05, fold change of greater than 2 and maximum CPM of greater than 1. Batch effect between different sequencing protocols was removed using removeBatchEffect of Limma package.

To investigate pathways enriched in sample groups, Gene Set Enrichment Analysis (GSEA)[8] was performed using FPKM (Fragments per Kilobase of transcript per Million mapped reads) values generated with RSEM[2] against mSigDB gene sets including hallmark, immunologic

signatures, motif, canonical pathways, gene ontology (GO) and oncogenic signatures. Network analysis for the differentially expressed genes was performed using STRING[9] v11 (Search Tool for the Retrieval of Interacting Genes/Proteins). Up-regulated genes ($P < 0.05$, fold change >2, count per million (CPM) >1) in responders (n=227 genes) and non-responders (n=118 genes) were analyzed separately. We used three STRING database association evidence channels as interaction sources to generate the gene-gene association: biochemical genetic data (experiments); previously curated pathway and protein-complex knowledge (databases); gene-by-gene correlation based on gene expression transcriptome/proteome data sets (co-expression). Minimum required interaction score was set to 0.4 and the disconnected nodes were removed from the final network. Gene Ontology (GO) enrichment analysis was performed using DAVID 6.8.

For long-read sequencing of primary ALL samples using the Oxford Nanopore platform, full-length cDNA was prepared from poly-A mRNA using Takara SMARTer cDNA synthesis. Multiplexed Nanopore libraries were prepared using the ligation sequencing kit (SQK-LSK109) with native barcoding (NBD104) and sequenced on GridION using MinKNOW v3.4.8 with real-time basecalling using Guppy v3.0.6 in high accuracy mode. To obtain a consensus sequence, all raw cDNA reads were aligned to the CD19.cAug10 isoform using minimap2 v2.17.

Whole exome/genome sequencing analysis

Paired-end WGS and WES reads were mapped to human reference genome GRCh37 by BWA[10] (version 0.7.12). Samtools[11] (version 1.4) was used to generate chromosomal coordinate-sorted and indexed bam files, and then processed by Picard (http://broadinstitute.github.io/picard/, version 2.0.1) MarkDuplicates module to mark PCR duplications. The reads were realigned around potential indel regions by GATK[12] (version 3.7) IndelRealigner module. Sequencing depth and coverage was assessed based on coding regions (~34Mb) defined by RefSeq genes. UnifiedGenotyper (within GATK v3.7) was applied to call SNVs and Indels from leukemia and

germline samples. The raw mutations were filtered by a homemade pipeline to exclude: 1) reported common SNPs/Indels from dbSNP v142; 2) germline mutations detected from matched germline control samples if available; 3) germline mutations observed in more than 2 samples within an in-house collection of germline cohort. The remaining variants were annotated using vep (v93)[13] and variants with maximum allele frequency of 1% or above were removed. All the non-silent SNVs/indels that passed the filtering pipeline were manually reviewed and only the highly reliable somatic ones were reported. Adjacent nucleotide changes observed on the same allele were merged into one mutation.

Copy number analysis using whole genome sequencing

Chromosome level copy numbers were estimated using coverage analysis of WGS. Control-FREEC (Control-Free Copy number caller)[14] software was used to estimate genome-wide copy-numbers using the mode without control sample. Read counts were corrected by GC content and mappability. Window size was automatically adjusted using coefficientOfVariation value of 0.08. Ploidy number was adjusted based on chromosome level copy numbers. Boundaries of focal copy number alterations (CNA) were modified if they overlapped with a structure variant call from Delly (v0.7.7).[15] Recurrent CNAs were analyzed using GISTIC2.[16]

Single cell sequencing and data analysis

Cryopreserved BMMC or PBMC samples were thawed, labeled and separated into three populations by FACS: tumor (CD45-dim/CD19+), T cells (CD45+/CD3+) and non-tumor non-T cells. Library preparation was performed using the Chromium Single Cell V(D)J Reagent Kit as per manufacturer's instructions (10x Genomics). 10,000 T cells per sample were used for T cell receptor (TCR) and 5′ Gene Expression (GEX) library preparation. 5,000 tumor cells and 5,000 non-tumor non-T cells (named "tumor mix") from each patient were pooled and used for and 5′ GEX library preparation. Libraries were sequenced on the NovaSeq 6000 System (Illumina): TCR libraries were sequenced to a target depth of 50-100 million paired-end 150bp reads per sample

(5,000-10,000 reads per cell); 5′ GEX libraries were sequenced to a target depth of 500 million paired-end 100bp reads per sample (50,000 reads per cell). The Cell Ranger Single-Cell Software Suite (version 3.0.0) was used to perform sample demultiplexing, barcode processing, single-cell gene counting, and TCR assembly (10x Genomics). Raw base call (BCL) files were demultiplexed using the Cell Ranger *mkfastq* pipeline into sample-specific FASTQ files that were then processed individually using the Cell Ranger *count* pipeline with default setting. Each sample was first analyzed individually using Seurat package (version 3.1.0) in R to select cells and to assign cell type to the selected cells[17, 18]. Paired αβTCRs associated with cells with gene expression information were annotated and characterized using TCRdist[19]. Paired CDR3s were compared to a subset of the VDJdb to assess putative specificity[20].

The initial Seurat object was generated by selecting cells with at least 200 genes detected and genes detected in at least 3 cells. Only cells with mitochondria content of less than 15% and hemoglobin content of less than 25% were kept for further analyses. The count data of selected cells was scaled to 10000 molecules per cell and the log2 transformed data was then scaled (z-score transformation). Most variable genes were selected using “mean.var.plot” method and were used for further single cell dimension reductions and clustering analysis. Cell type for each tSNE cluster was assigned based on expression of known cell markers. To compare responder and non-responder samples, four pre-blinatumomab samples were analyzed together. Anchor genes were identified for each pair of samples and were used to transform all samples into a shared space using canonical correlation analysis in Seurat. tSNE analysis was performed with the integrated set and differential expression analysis between responder and non-responder for each cluster as well as major cell clusters. Average expression for cells in the above-mentioned clusters and gene ranks was calculated, which was used for GSEA analysis. Known mutation data from genomic analysis was checked in the single cell data using cb_sniffer.[21]

CD19 exon junction analysis

Splicing junctions from STAR mapping were generated by merging junctions based its chromosome location and intron motif type, and annotated using Ensemble (v75)[22], followed by GENCODE v31 (lifted to GRCh37)[23] and AceView[24] gene structures. Three samples with low CD19 expression were excluded from the analysis as the exon junction usage could not be accurately estimated due to low number of reads. Splicing junctions with CPM value greater than 0.5 in more than 5 samples were kept for further analysis. The raw read counts were normalized using Limma/Voom. Normalized CD19 exon junctions were then extracted at locus chr16:28943259-28950668. A total of 18 junctions were identified, including 15 exon junctions from 2 CD19 isoforms (NM_001770, NM_001178098), two junctions with either exon 2 or exon 12 skipping, and one junction with partial exon 2 deletion. Some reads mapped to the non-canonical junctions were soft clipped. After mapping of these soft-clipped sequences using BLAT[25], the soft clipped reads mapped to the exon 2-3 junction (EJ_{2-3}). Reads mapped to CD19 locus were extracted and aligned using BLAT against chr16, and exon junctions were generated by parsing the psl file.

To confirm the presence of *CD19* exon 2 skipping and partial exon 2 deletion isoforms, primers were designed to amplify exon 1 to exon 4 of *CD19* (forward 5'GCCCCGGAGAGTCTGACCACCATGC, reverse 5'CCCACATATCTCTGGCCGGGCGATCG). Total RNA was extracted from KOPN75 cell line and cDNA was prepared by using SuperScript III First-Strand Synthesis System (Thermo Fisher). cDNA samples were amplified by PCR using Advantage 2 Polymerase Mix (Takara) and *CD19* isoforms were visualized in 1.5% agarose gels. For fragment size analysis, the forward primer was conjugated with FAM at its 5' end, and the amplified PCR product was column purified (Promega) and analyzed by capillary electrophoresis on a 3500xl Genetic Analyzer (Life Technologies). To confirm the full-length transcripts of *CD19*, the coding sequence was PCR amplified (forward primer 5' AGTCTGACCACCATGCCACCTCCTC, reverse primer

5‘ACACACACTTACACACATGCACACA), gel purified and subject to TA cloning using the pGEM-T Easy Vector System II (Promega). Sanger sequencing was performed on selected colonies and aligned to the *CD19* WT reference sequence in CLC Main Workbench 20 (Qiagen).

Targeted sequencing and cloning of CD19

Five pairs of primers were designed to amplify exon 2, 3 and 4 of *CD19* (supplementary Table 1). The targeted sequence was amplified using Phusion High-Fidelity DNA Polymerase (New England Biolabs). PCR products were purified using Wizard SV Gel and PCR Clean-Up System (Promega). Sequencing libraries were generated by Nextera XT transposase-based library preparation using KAPA HyperPrep Kits (Illumina) and sequenced using MiSeq Reagent 500-cycles Nano Kit v2 (Illumina) on a MiSeq sequencer to 20,000x average coverage.

To generate *CD19* mutant 1 (p.Tyr259fs, chr16 28944770 T→TGT) and *CD19* mutant 2 (p.Tyr259fs, chr16 28944771 A→ATTGGAGATCCC), insertions were introduced into a MSCV-CD19-IRES-RFP retroviral vector using Q5 Site-Directed Mutagenesis Kit (New England Biolabs). All insertions were validated by Sanger sequencing. To produce ecotropic retrovirus, HEK-293T cells were co-transfected with CAG4-Eco and pMD-MLV-ogp helper packaging plasmids using FuGENE HD (Promega) and MSCV-IRES-RFP empty vector, WT or mutant *CD19*. Viral supernatants were used to infect NIH-3T3 mouse fibroblast cells in the presence of polybrene (10 µg/mL).

Flow cytometry and Immunofluorescence

Transduced NIH-3T3 cells were dissociated using TrypLE Express Enzyme (Gibco, Thermo Fisher Scientific), and stained with CD19-FITC (clone HIB19, BD Biosciences). Transduced cells were gated on RFP+ and the expression level of CD19 was assessed. For immunofluorescence staining, NIH-3T3 cells expressing WT or mutant CD19 were seeded overnight to poly-D-lysine-coated Millicell EZ slides (Millipore), fixed for 10 min at room temperature with 4% paraformaldehyde and washed three times. Sites of nonspecific antibody

binding were blocked by incubating cells for 60 min in goat serum (Sigma-Aldrich) 1× /PBS. Cells were stained at room temperature for 1 h with anti-CD19 (clone HIB19, Santa Cruz Biotechnology), washed and then incubated for 45 min with a secondary antibody conjugated to Goat anti-mouse IgG (H+L), Alexa Fluor 488 (Thermo Fisher Scientific, A-28175). Slides were washed and were mounted with Golden ProLong Diamond Antifade Mountant with DAPI (Life Technologies), and fluorescent images were captured using Nikon C2confocal microscope.

Supplemental Table 1. Oligonucleotides for CD19 targeted sequencing

Primer name	Amplicon	Sequence (5'-3')
CD19_exon2_F1	CD19 exon 2	cacccatgcccacacctctctcC
CD19_exon2_R1	CD19 exon 2	GCCCCGGCTGGCACAGGTAGAAG
CD19_exon2_F2	CD19 exon 2	CCCTGGCCATCTGGCTTTTCATCTT
CD19_exon2_R2	CD19 exon 2	TAGCTCCGAGTTTCGCCCTCAATCC
CD19_exon3_F	CD19 exon 3	TGGGTGTCTCTGCATTTGGTTCTGG
CD19_exon3_R	CD19 exon 3	CTTCCCGGCATCTCCCCAGTCAT
CD19_exon4_F1	CD19 exon 4	GGGGAAACTGAAGAGGTGAAACCCTGA
CD19_exon4_R1	CD19 exon 4	CAACAGACCCGTCTCCATTACCCACA
CD19_exon4_F2	CD19 exon 4	ACCCCAAGGGGCCTAAGTCATTGCT
CD19_exon4_R2	CD19 exon 4	CTTCCCAGTACCCCCACACAGATGC

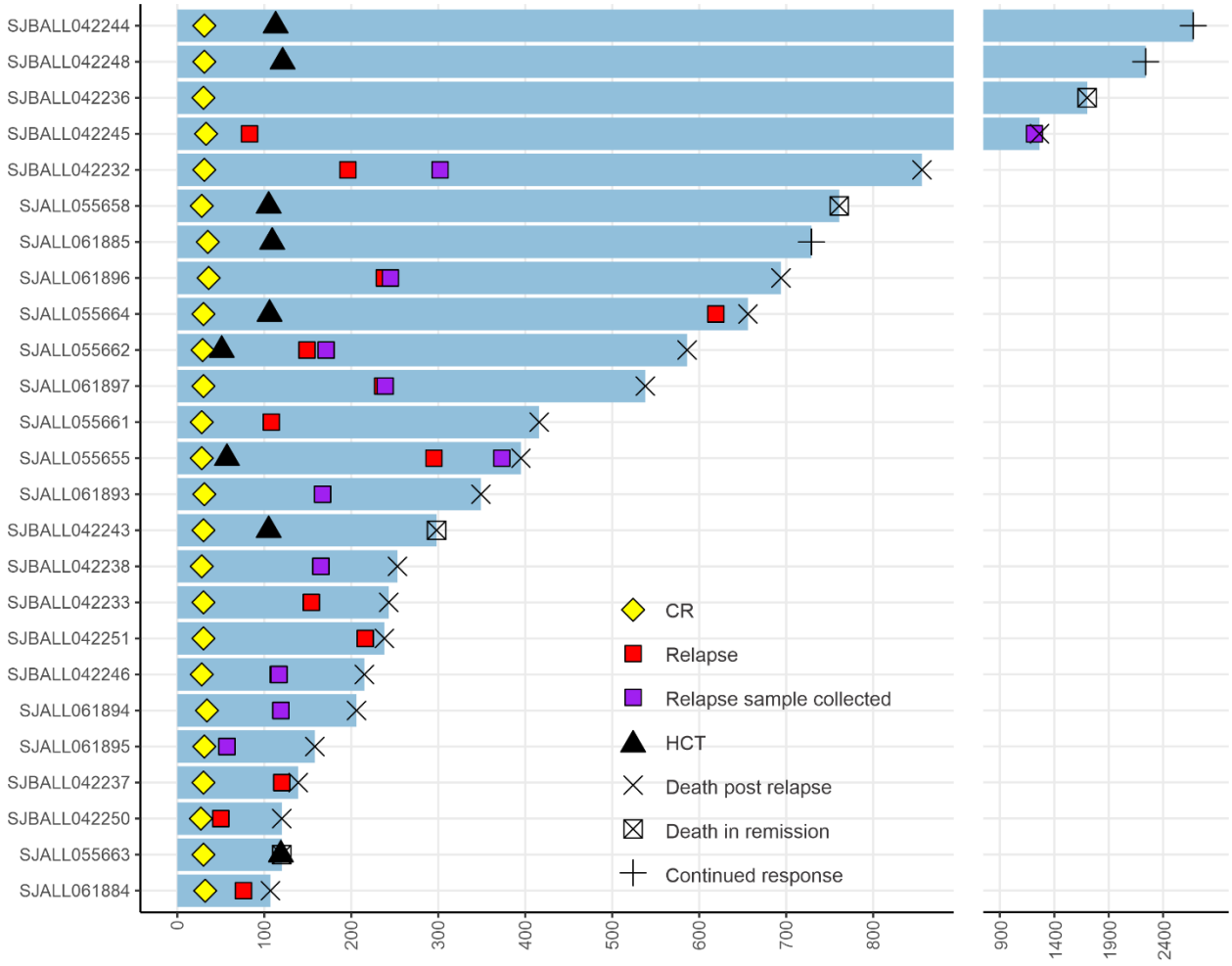
Supplemental Table 2. Cohort of blinatumomab treated r/r B-ALL
Supplemental Table 3. Output of gene fusions
Supplemental Table 4. Output of SNVs/indels
Supplemental Table 5. Differentially expressed genes in responders and non-responders
Supplemental Table 6. GO biological processes in responders
Provided as Excel tables in "Supplemental Tables".

Supplemental Table 7. GSEA results of responders vs non-responders

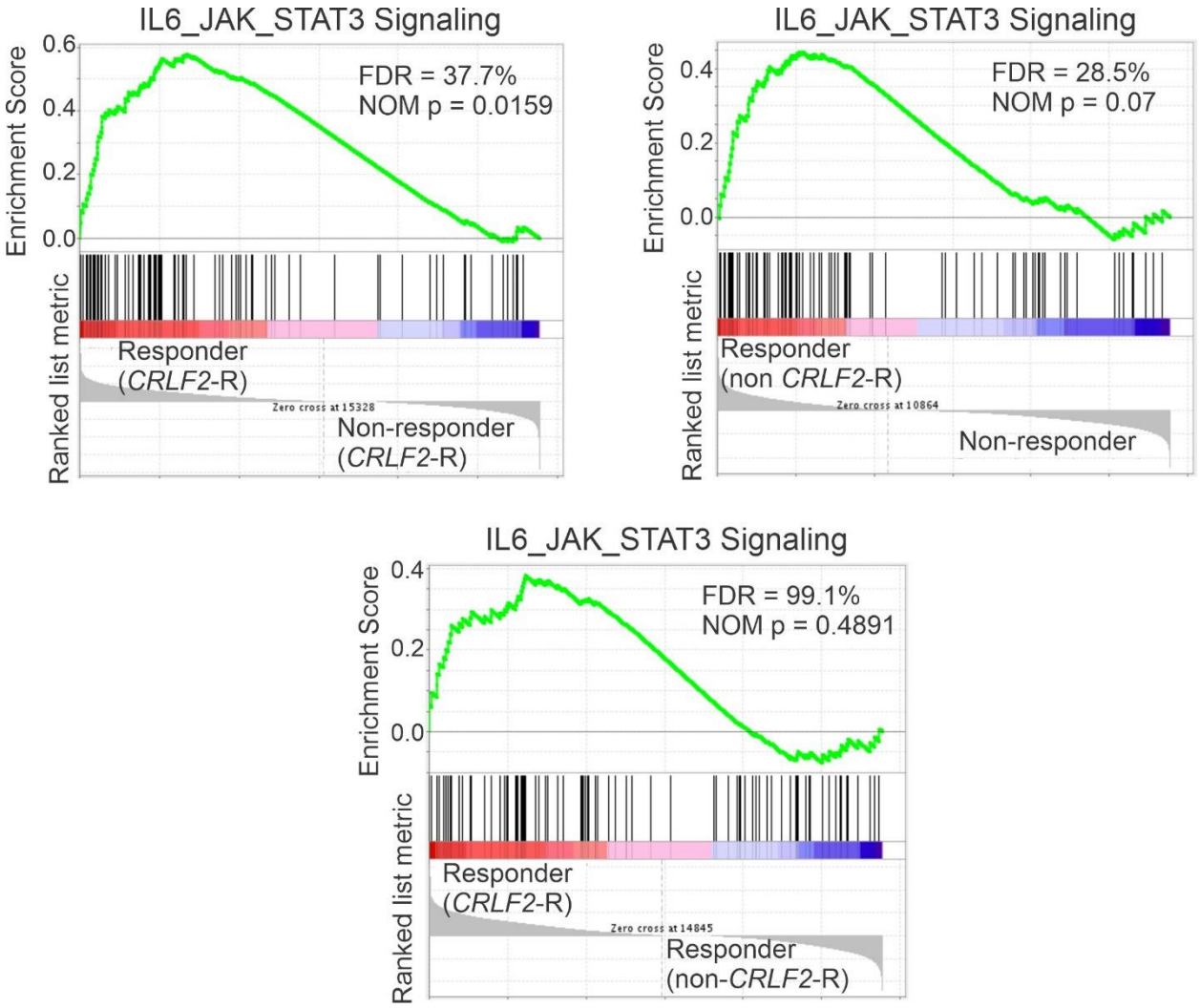
Name	NOM p-val	FDR q-val
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.0000	0.0196
HALLMARK_HYPOXIA	0.0021	0.1119
HALLMARK_IL6_JAK_STAT3_SIGNALINGp	0.0044	0.1577
HALLMARK_ALLOGRAFT_REJECTION	0.0152	0.1341
HALLMARK_INFLAMMATORY_RESPONSE	0.0022	0.1122
HALLMARK_TGF_BETA_SIGNALING	0.0299	0.1745
HALLMARK_APOPTOSIS	0.0285	0.1754
HALLMARK_IL2_STAT5_SIGNALING	0.0154	0.1764
HALLMARK_HEME_METABOLISM	0.1152	0.1807
HALLMARK_COMPLEMENT	0.0413	0.1908
HALLMARK_ESTROGEN_RESPONSE_EARLY	0.0388	0.2435
HALLMARK_P53_PATHWAY	0.0802	0.2479

Supplemental Table 8. CD19 targeted sequencing
Supplemental Table 9. CD19 exon 2 partial deletion usage
Provided as Excel tables in "Supplemental Tables".

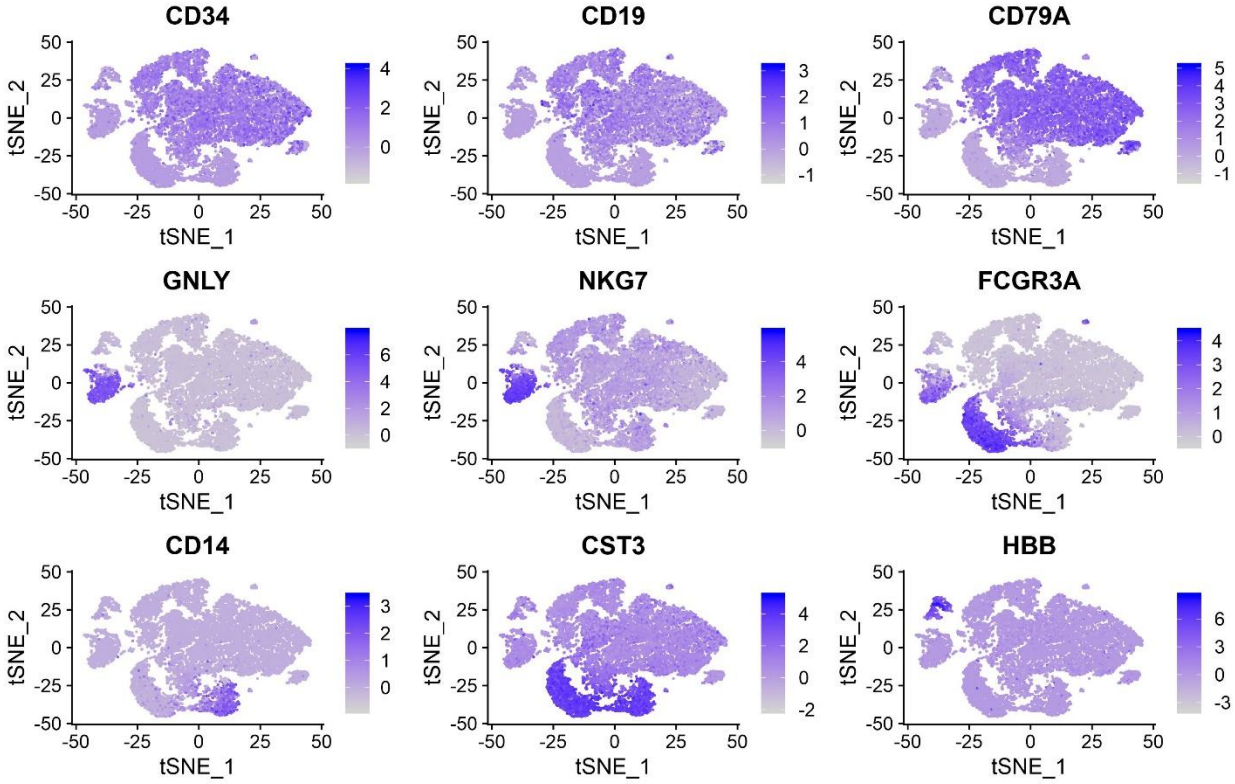
Supplemental Figure 1. Swimmer plot of responders to blinatumomab (n=25) in this cohort.



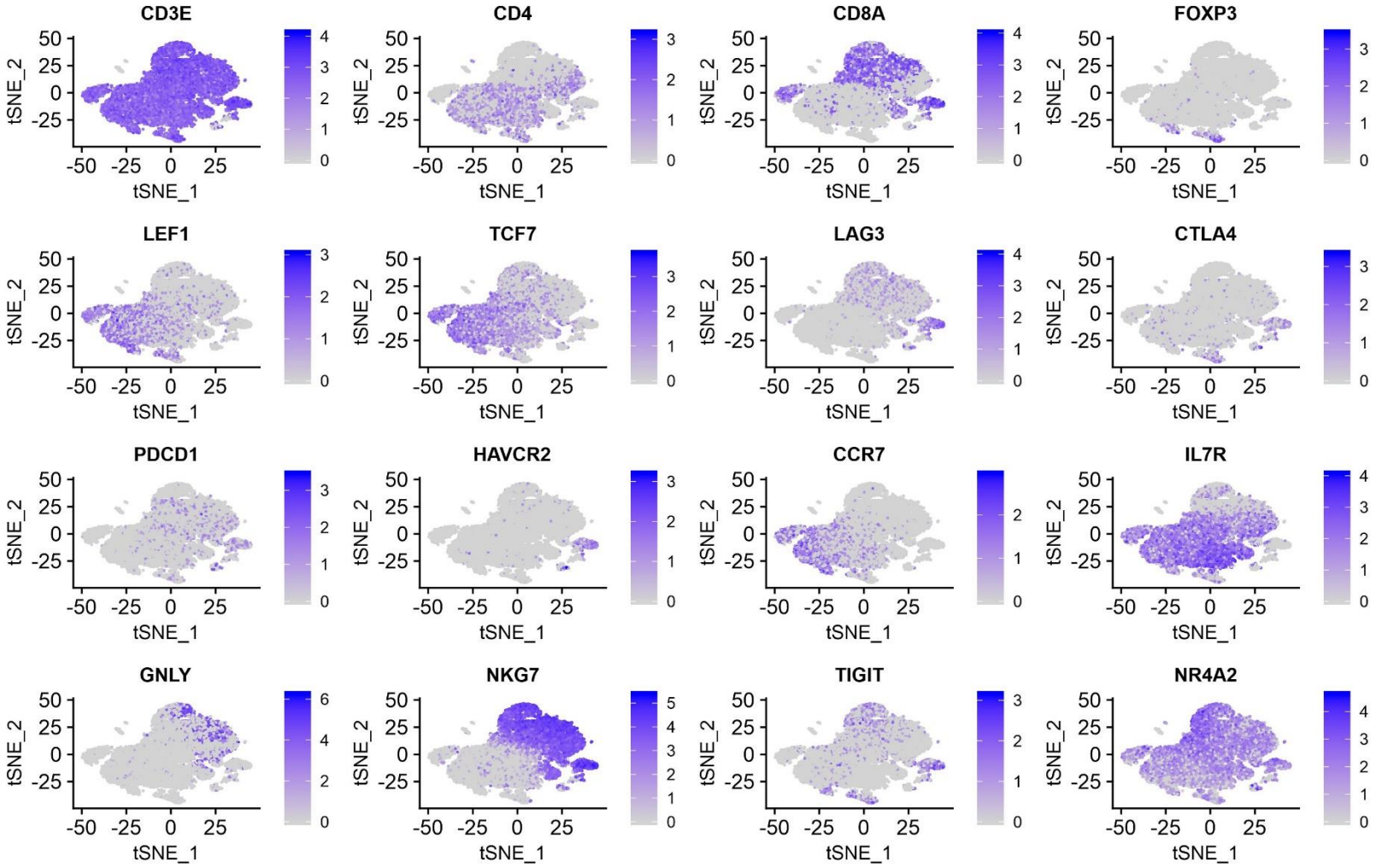
Supplemental Figure 2. GSEA for IL6-JAK-STAT3 signaling pathway demonstrating JAK-STAT activation in responders is independent of CRLF2-R.



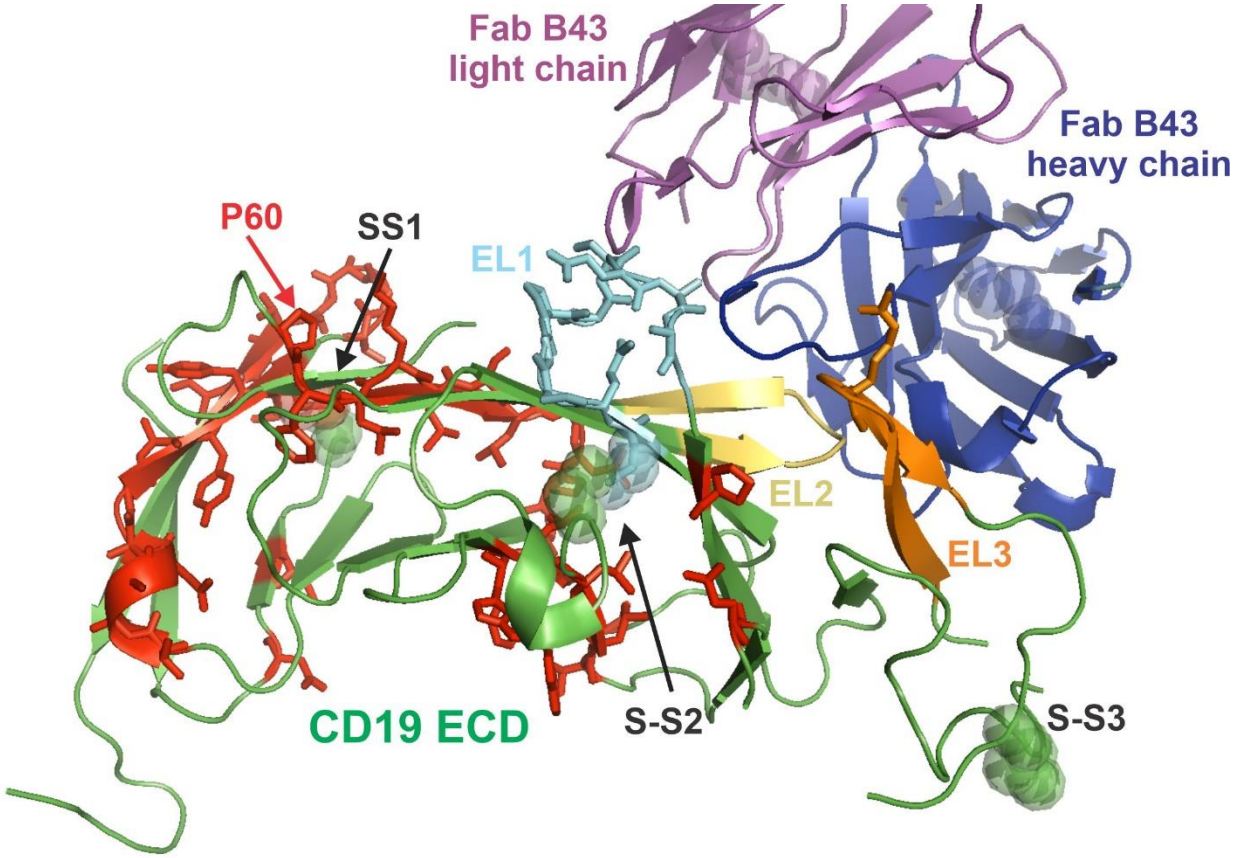
Supplemental Figure 3. 10x Genomics 5'GEX tSNE plots of tumor mix cells (CD19+ blasts and non-tumor, non-T cells) from two blinatumomab responders and two non-responders. Tumor cells are positive for CD34, CD19 and CD79A. NK cells are positive for GNLY and NKG7. CD16+ monocytes are positive for FCGR3A and CST3. CD14+ monocytes are positive for CD14 and CST3.



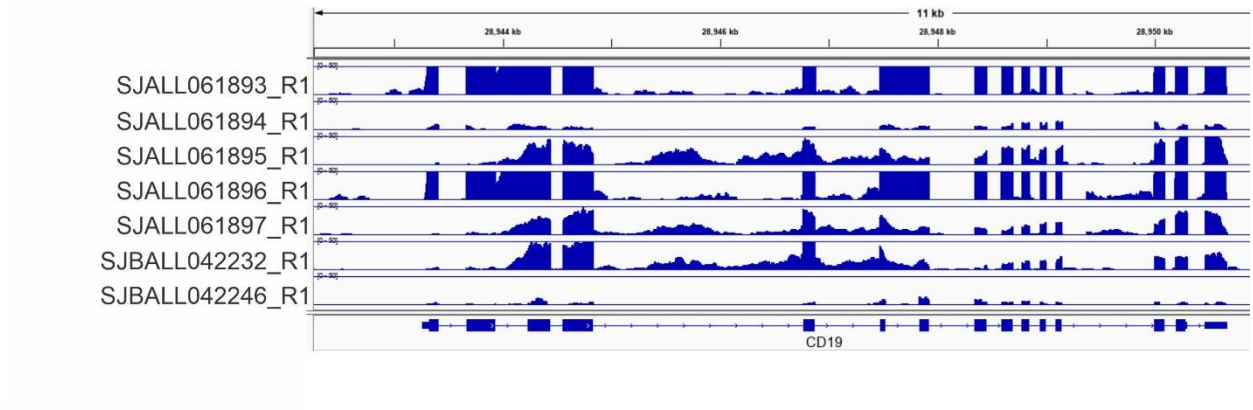
Supplemental Figure 4. 10x Genomics gene expression tSNE plots of CD3+ T cells from two blinatumomab responders and two non-responders. Expression of genes that distinguish different T cell subsets.



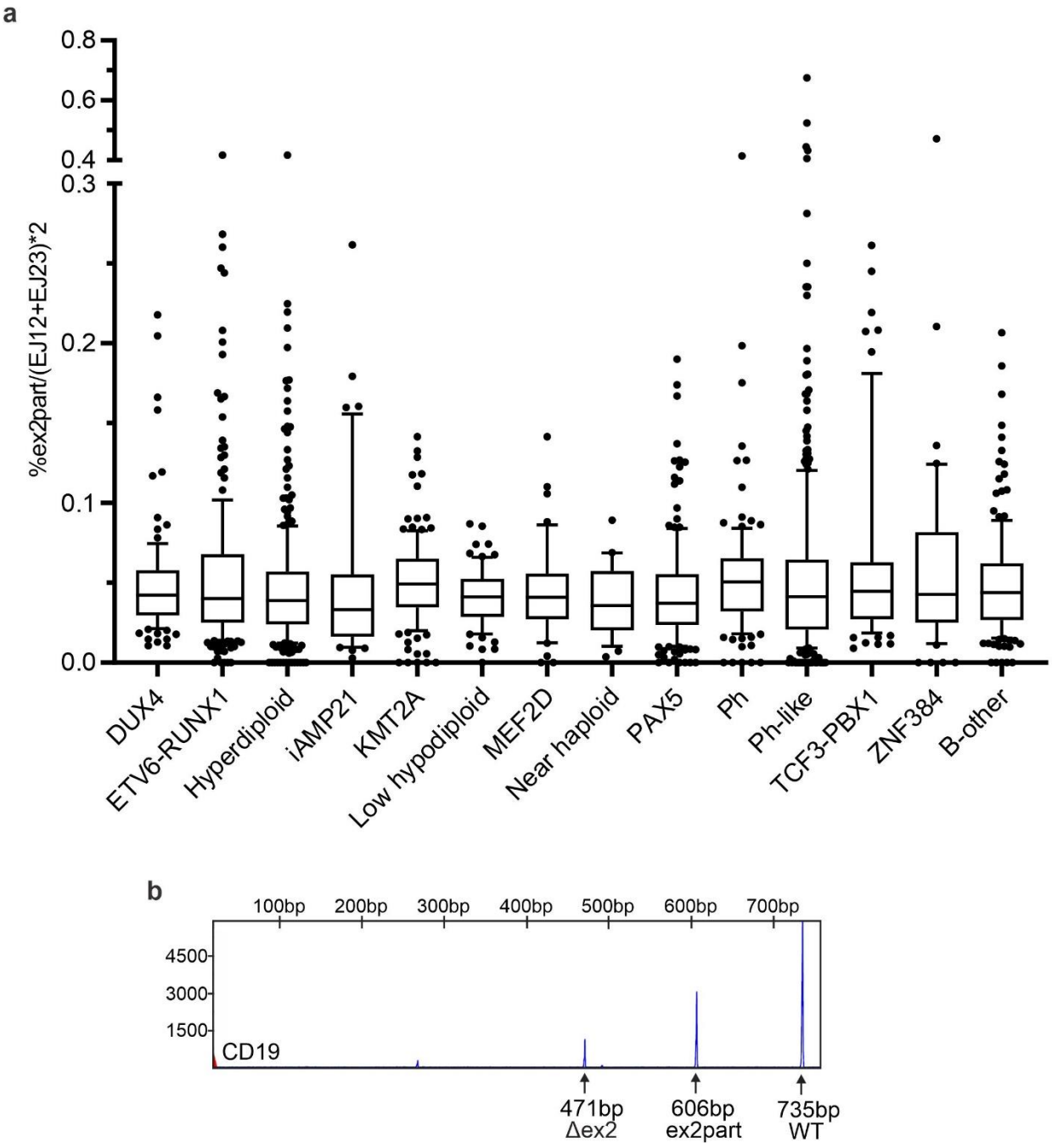
Supplemental Figure 5. Crystal structure of CD19 extracellular domain (ECD) with B43 Fab (pdb: *6al5*), the CD19 recognition arm in blinatumomab. Spheres represent di-sulfide bonds: SS1 (Cys38-Cys261); SS2 (Cys97-Cys200); SS3 (Cys134-Cys173). Three epitope loops: EL1 (97-107); EL2 (155-166); EL3 (216-224). Sites of mutation are highlighted red and shown as sticks. Mutations were observed in two of the epitope loops and the core CD19 structure, which interferes with protein stability.



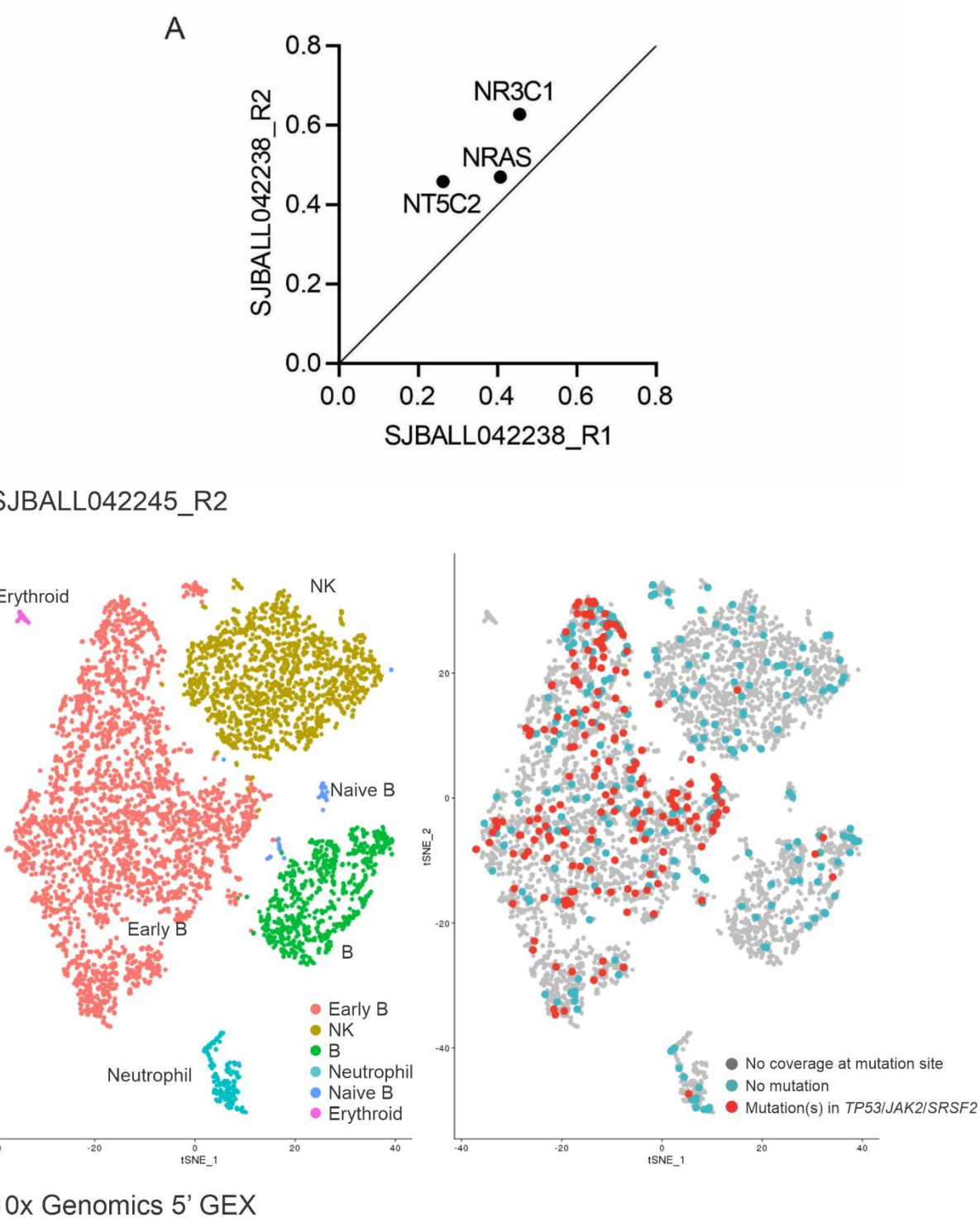
Supplemental Figure 6. Expression levels of CD19 by RNA-seq.



Supplemental Figure 7. (A) Levels of ex2part in B-ALL cohort of diagnosis samples (n=1988).[26] We evaluated the relative usage of ex2part by calculating the percentage of ex2part junction compared to the average of the canonical exon junctions EJ_{1-2} and EJ_{2-3} (% (ex2part / EJ_{1-2} + EJ_{2-3})*2; median usage of 4%). (B) Fragment size analysis of CD19 showing three isoforms: wild-type (WT), ex2part and ex2skip (Δex2).



Supplemental Figure 8. (A) Mutations in pre-blinatumomab (SJBALL042238_R1) and post-blinatumomab (SJBALL042238_R2) identified by whole exome sequencing. (B) 10x Genomics 5' GEX of tumor mix cells from post-blinatumomab sample SJBALL042245_R2. Expression clusters are shown on the left and cell with detectable mutation are highlighted on the right, showing localization of mutations in the predominant CD19+ blast population.



References

1. A Dobin, CA Davis, F Schlesinger, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15-21.
2. B Li and CN Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12(323.
3. R Schmieder and R Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27(6):863-864.
4. D Nicorici, M Satalan, H Edgren and S Kangaspeska. FusionCatcher-a tool for finding somatic fusion genes in paired-end RNA-sequencing data. . *bioRxiv*. 2014;011650(
5. JT Robinson, H Thorvaldsdottir, W Winckler, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;29(1):24-26.
6. NL Bray, H Pimentel, P Melsted and L Pachter. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525-527.
7. CW Law, Y Chen, W Shi and GK Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15(2):R29.
8. A Subramanian, P Tamayo, VK Mootha, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550.
9. D Szklarczyk, AL Gable, D Lyon, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613.
10. H Li and R Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
11. H Li, B Handsaker, A Wysoker, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
12. MA DePristo, E Banks, R Poplin, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498.
13. W McLaren, L Gil, SE Hunt, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
14. V Boeva, T Popova, K Bleakley, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. 2012;28(3):423-425.
15. T Rausch, T Zichner, A Schlattl, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*. 2012;28(18):i333-i339.
16. CH Mermel, SE Schumacher, B Hill, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011;12(4):R41.
17. A Butler, P Hoffman, P Smibert, E Papalexi and R Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36(5):411-420.
18. T Stuart, A Butler, P Hoffman, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902 e1821.
19. P Dash, AJ Fiore-Gartland, T Hertz, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*. 2017;547(7661):89-93.
20. DV Bagaev, RMA Vroomans, J Samir, et al. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res*. 2020;48(D1):D1057-D1062.
21. AA Petti, SR Williams, CA Miller, et al. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat Commun*. 2019;10(1):3660.

22. T Hubbard, D Barker, E Birney, et al. The Ensembl genome database project. *Nucleic Acids Res*. 2002;30(1):38-41.
23. J Harrow, A Frankish, JM Gonzalez, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22(9):1760-1774.
24. D Thierry-Mieg and J Thierry-Mieg. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol*. 2006;7 Suppl 1(S12.11-14.
25. WJ Kent. BLAT--the BLAST-like alignment tool. *Genome Res*. 2002;12(4):656-664.
26. Z Gu, ML Churchman, KG Roberts, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat Genet*. 2019;51(2):296-307.