SUPPLEMENTAL MATERIAL

Profiling chromatin accessibility in pediatric acute lymphoblastic leukemia identifies subtype-specific chromatin landscapes and gene regulatory networks

Jonathan D. Diedrich[1,2], Qian Dong[1,2], Daniel C. Ferguson[1,2], Brennan P. Bergeron[1,2,8], Robert J. Autry[1,2,3], Maoxiang Qian[1,2], Wenjian Yang[1,2], Colton Smith[1,2], James B. Papizan[6], Jon P. Connelly[6], Kohei Hagiwara[5], Kristine R. Crews[1,2], Shondra M. Pruett-Miller[6], Ching-Hon Pui[1,4,7], Jun J. Yang[1,2,4], Mary V. Relling[1,2], William E. Evans[1,2] and Daniel Savic[1,2,3,9]

[1] Hematological Malignancies Program and Center for Precision Medicine in Leukemia, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.

[2]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.

[3]Integrated Biomedical Sciences Program, University of Tennessee Health Science Center, Memphis, TN 38105, USA.

[4]Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.

[5]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.

[6]Department of Cell and Molecular biology and Center for Advanced Genome Engineering, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.

[7]Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.

[8]Graduate School of Biomedical Sciences, St. Jude Children's Research Hospital, Memphis, TN 38105, USA.

[9]Corresponding author: Daniel Savic; daniel.savic@stjude.org

SUPPLEMENTAL METHODS

Selection and Preparation of cryopreserved primary ALL cells from St. Jude patients

Cases were selected based on biological and technical considerations because of limited access to cryopreserved primary ALL cell biospecimens. We focused on 3 common ALL subtypes that are driven by distinct types of malignant driver events (aneuploidy= hyperdiploid, gene fusion= ETV6-RUNX1 and complex rearrangement= DUX4/ERG). This allowed our results to be extensible to a large number of ALL cases and further allowed us to assess how distinct malignant transformation events impact open chromatin accessibility. In addition, we were limited by the total number of vials available in cryo (i.e. each sample needed to have >=2 cryopreserved vials to ensure the these same biospecimens was available for additional/future leukemia studies at St. Jude), as well as the total number of cryopreserved cells per vial (i.e. >1 million cells to ensure enough viable cells/volume after cell sorting for technical replicate Fast-ATAC experiments and/or >5 million cells to ensure enough viable cells for ATAC-seq and RNA-seq experimentation). Cryopreserved samples were thawed using pre-warmed high glucose Dulbecco's Modified Eagle's Medium (DMEM) containing 60% fetal bovine serum (FBS). Cells were then spun down at 300xg for 10 minutes at 4°C. After centrifugation, the cell pellet was resuspended in 1 mL of high glucose DMEM containing 20% FBS. To remove dead primary ALL cells and to ensure only viable cells are used as input for functional genomic assays, cells were flow sorted for viable cells using DAPI staining at the St. Jude Children's Research Hospital Flow Cytometry and Cell Sorting Core.

ATAC-seq in primary ALL cells

We performed open chromatin mapping using the ATAC-seq FAST-ATAC protocol as previously reported (1). We used 10,000 cells for ATAC-seq and all samples were sequenced on the Illumina HiSeq next-generation sequencing (NGS) platform at the Hartwell Center for Bioinformatics and Biotechnology at St. Jude Children's Research Hospital. We used paired-end 100-basepair NGS and sequencing reads were mapped to the human hg19 reference genome using bowtie2 (2). After aligning the sequencing reads to the genome, we removed PCR duplicates and reads that mapped to mitochondrial DNA, and identified open chromatin sites using the MACS2 peak caller (3) by using the BAMPE setting and a default q value of 0.05. We identified high-confidence open chromatin sites that were reproducibly identified in two or more primary ALL cell samples for analyses comparing accessible chromatin between ALL cells and B-cells, or in each ALL subtype for analyses comparing accessible chromatin between ALL subtypes, and we merged all reproducible open chromatin sites using bedtools (4). To identify enriched open chromatin sites, we assessed for differences in open chromatin accessibility using DESeq2 (5) on NGS read depth at the union of all reproducible open chromatin sites. Sex and batch were used as covariates for DESeq2 (5) analyses. Principal component analysis (PCA) using normalized read depth at open chromatin sites was performed using the *prcomp* function in R. Spearman rank correlations of normalized read depth at the union of open chromatin sites was calculated from pair-wise comparisons, and this was used as input for unsupervised hierarchical clustering using the *heatmap.2* function from gplots v3.0.1 in R. HINT-ATAC (6) was used to map TF footprints and to generate TF activity scores. To identify TF footprints in each ALL subtype using HINT-ATAC, we merged BAM files for samples in each subtype and used open chromatin sites that were reproducibly identified in each ALL subtype. Patient ATAC-seq data has been deposited to NCBI Gene Expression Omnibus (GSE161501).

<u>RNA-seq in primary ALL cells</u>

Stranded RNA-seq on primary ALL cell samples from patients were performed by the Hartwell Center for Bioinformatics and Biotechnology at St. Jude Children's Research Hospital on the Illumina HiSeq platform. Total RNA was purified from 19 of 24 primary ALL cell samples using Norgen Total RNA Purification Kits (ETV6-RUNX1= 6, DUX4/ERG= 2, Hyperdiploid= 11). We obtained fragments per kilobase of transcript per million mapped reads (FPKM) and raw NGS read counts for all genes using the St. Jude WARDEN pipeline on DNAnexus (https://www.dnanexus.com). Gene counts for *DUX4* were identified as previously described (7, 8). DESeq2 (5) was used to identify DEGs among ALL subtypes using NGS read gene counts. PCA using normalized gene counts was performed using the *prcomp* function in R. Spearman rank correlations of normalized gene counts were also calculated from pair-wise comparisons of primary cell samples and this was used as input for unsupervised hierarchical clustering using the *heatmap.2* function from gplots v3.0.1 in R. Gene set enrichment analysis (GSEA) was used to identify biological pathways of DEGs associated with differentially accessible sites identified between ALL cells and normal B-cells (9). We generated cumulative distribution functions (CDFs) in R using the *ecdf* function. To generate CDFs, we compared the distances of DEG transcription start sites (TSS) to the nearest differentially accessible open chromatin site. As background, we computed the distances of TSS of all expressed genes in each subtype to the nearest differentially accessible open chromatin site. To identify expressed genes in each ALL subtype, we used NGS read gene count cutoffs of >20 for all cell samples within an ALL subtype. Patient RNA-seq data has been deposited to NCBI Gene Expression Omnibus (GSE161501). Stranded RNA-seq on WT/parental (n=3) and *TCFL5* KO (n=3) REH cells were also performed by the Hartwell Center for Bioinformatics and Biotechnology at St. Jude Children's Research Hospital on the Illumina HiSeq platform. Total RNA was purified from REH cells using Norgen Total RNA Purification Kits. The St.

Jude WARDEN pipeline was used to obtain gene FPKM and to identify differentially expressed genes between WT and *TCFL5* KO REH cells.

ChIP-seq in Nalm6 cells

ChIP-seq was performed as previously described (10). Briefly, 20 million Nalm6 were crosslinked using 1% formaldehyde for 10min and sonicated on a Diagenode Bioruptor Plus instrument. Chromatin immunoprecipitation was performed using 5μg anti-DUX4 antibody (Abcam, ab124699). Samples were run on an Illumina NovaSeq next-generation sequencing (NGS) machine using single-end 100bp sequencing. Following NGS, reads were mapped to the hg19 reference genome using BWA (11) and binding sites were called using MACS2 peak caller (3) using a default q value of 0.05.

Publicly available functional genomic datasets

Raw ATAC-seq and RNA-seq data from Corces *et al.* (1) and from Calderon *et al.* (12) was downloaded from the NCBI Gene Expression Omnibus for normal hematopoietic cells (GSE74912 and GSE118189), and analyzed in an identical manner as primary ALL cell samples. For comparisons of ATAC-seq between B-cells and ALL cells, high-confidence open chromatin sites that were reproducibly identified in two or more B-cells were used for analysis, and we assessed for differences in open chromatin accessibility using DESeq2 (5). PCA using normalized read depth at open chromatin sites was performed using the *prcomp* function in R. Spearman rank correlations of normalized read depth at the union of open chromatin sites was calculated from pair-wise comparisons, and this was used as input for unsupervised hierarchical clustering using the *heatmap.2* function from gplots v3.0.1 in R. To measure and identify differences in gene expression between B-cells and ALL cells we used DESeq2 (5). ChromHMM data (13, 14) from GM12878 B-cell lymphoblastiod cell line was downloaded from the UCSC genome browser (https://genome.ucsc.edu/). Chromatin ChIP-seq data for primary B-cells was downloaded from the Blueprint Epigenome consortium (https://www.blueprint-epigenome.eu/). Publicly available DNA methylation data from Nordlund *et al.* (15, 16) was downloaded from NCBI GEO (GSE49031).

ATAC-seq in human ALL cell lines

The ATAC-seq FAST-ATAC protocol (1) was performed on 697, Nalm6, REH, SEM, SUPB15 and UOCB1 human ALL cell lines and analyzed as described above, and these data can be found on the NCBI Gene Expression Omnibus (GSE129066).

DNA methylation analyses

Illumina Infinium HumanMethylation450K BeadChIP CpG DNA methylation array data (17) from Total Therapy XVI (TOTXVI) were available for 19 of the 24 primary ALL cells analyzed in this study (ETV6-RUNX1= 4, DUX4/ERG= 7, Hyperdiploid= 8), and these data were obtained from NCBI Gene Expression Omnibus (GSE66708). We determined DNA methylation beta values at all CpG sites and performed PCA of DNA methylation beta values using the *prcomp* function in R. To determine DNA methylation beta-values at differentially accessible open chromatin sites, we utilized bedtools (4) to identify DNA methylation probes that mapped to differentially accessible open chromatin sites. For each DNA methylation probe, we calculated the median DNA methylation beta-value across primary ALL cells samples within each ALL subtype or within pooled ALL cell samples from opposing subtypes. To test for significant differences in DNA methylation at differentially accessible open chromatin sites, we performed Wilcoxon rank-sum tests on median DNA methylation beta values between ALL subtypes and their two opposing subtypes.

Gene regulatory network analysis

Gene regulatory networks for 19 ALL patient samples with both ATAC-seq and RNA-seq data was generated by PECA (18) using ATAC-seq BAM files and RNA-seq TPM (transcripts per million) count files. “TF-target gene” connections across all samples within each ALL subtype were combined and redundant connections were removed, and the number of connections to each target gene within each ALL subtype were calculated in Linux. We compared network target genes between subtypes by subtracting the number of target gene connections between two subtypes and ranking them. We set 75 network connections as the threshold for enriched target genes in each subtype compared to opposing subtypes. The gene regulatory network maps of enriched target genes was generated by Cytoscape (19) and the nodes with less than 150 neighbors within distance one of enriched target genes were filtered.

CRISPR/Cas9 genome editing

TCFL5 knockdown pools were generated using CRISPR-Cas9 technology. Briefly, 400,000 REH cells were transiently transfected with precomplexed ribonuclear proteins (RNPs) consisting of 100pmol of chemically modified sgRNA (5’ – AAGCAUUUGUAGUAAACAGU- 3’, Synthego) and 35pmol of Cas9 protein (St. Jude Protein Production Core) via nucleofection (Lonza, 4D-Nucleofector™ X-unit) using solution P3 and program CA-137 in a small (20ul) cuvette according to the manufacturer’s recommended protocol. Targeted amplicons were generated using gene specific primers with partial Illumina adapter overhangs (hTCFL5.F – 5’-AGCCAGAGCCATGGGAGTGGGATGG-3’ and hTCFL5.R

– 5’-GCCTTGGCGCCCGGCTTAAAAGGTT-3’, overhangs not shown) and sequenced as previously described (20). Briefly, cell pellets of approximately 10,000 cells were lysed and used to generate gene specific amplicons with partial Illumina adapters in PCR#1. Amplicons were indexed in PCR#2 and pooled with targeted amplicons from other loci to create sequence diversity. Additionally, 10% PhiX Sequencing Control V3 (Illumina) was added to the pooled amplicon library prior to running the sample on an Miseq Sequencer System (Illumina) to generate paired 2 X 250bp reads. Samples were demultiplexed using the index sequences, fastq files were generated, and NGS analysis was performed using CRIS.py (21).

DSC3, *IGF2BP1* and *KCNN1* enhancer deletion REH cell pools were generated using CRISPR-Cas9 technology. In brief, one million REH cells were transiently transfected with precomplexed ribonuclear proteins (RNPs) consisting of 100pmol of each chemically modified sgRNA (Synthego), 35pmol of Cas9 protein (St. Jude Protein Production Core), and 3ug of ssODN (Alt-R modifications, IDT) via nucleofection (Lonza, 4D-Nucleofector™ X-unit) using solution P3 and program CA-137 in a small (100ul) cuvette according to the manufacturer’s recommended protocol. Three days post-nucleofection, genomic DNA was harvested via crude lysis and used for PCR amplification. The presence of the desired deletion was confirmed via gel electrophoresis and sequencing. Editing construct sequences and relevant primers are listed in the table below.

Name	**Sequence (5’ to 3’)**
KCNN1 reagents	
CAGE888.KCNN1.g7 sgRNA spacer	UGGAGGUGGGAACUGUGGCG
CAGE889.KCNN1.g1 sgRNA spacer	GCCAAGUUCAGCCUGGGUGC
CAGE888.g7.CAGE889.g1.sense.ssODN	GGGGAACAGCAAGTGCAAAGGCCTGGAGGTGGGAACTGTGGAATTCTGCTGGGGGAAGGAGCACAGTTTTGGGGACCCCCAGCCCT
CAGE888.KCNN1.F	AGCTGTCAGGGGCCAGGAGCATTGA
CAGE889.KCNN1.R	TGGGAGGAGAGACGCCCGTC
DSC3 reagents	
CAGE883.DSC3.g7 sgRNA spacer	AGUUUUCAGAAUUGUCCGUA
CAGE884.DSC3.g1 sgRNA spacer	UCUAUCAUCUACAUUAUGAG
CAGE883.g7.CAGE884.g1.sense.ssODN	GGATTTATTACCATTTATTAAAGAGTTTTCAGAATTGTCCGAATTCGAGAGGACACTGTAGAAGGAAAATGAAACAGATATTGAGC
CAGE883.DSC3.F	ACAGCCTCCCATCTCAATTAGCAGGG
CAGE884.DSC3.R	TCTCAATAAAATGCACCTATTCCAA

IGF2BP1 reagents	
CAGE879.IGF2BP1.g25 sgRNA spacer	GUCGGGUUUCGACCGGCCGG
CAGE880.IGF2BP1.g44 sgRNA spacer	AGGACACACGCUCAGGCACU
CAGE879.g25.CAGE880.g44.anti.ssODN	CACTGCACTCCAGCCTGGGCGACAGGACACACGCTCAGGCGAATTCGCCGGTCGAAACCCGACCCCATGGCGAAGCCAGGCAGCCG
CAGE879.IGF2BP1.F	CCTGGTGCAGGCGGGAAGC
CAGE880.IGF2BP1.R	GGCCAACATGGTGAAACCCCGTCTC

DNA sequence variant analyses

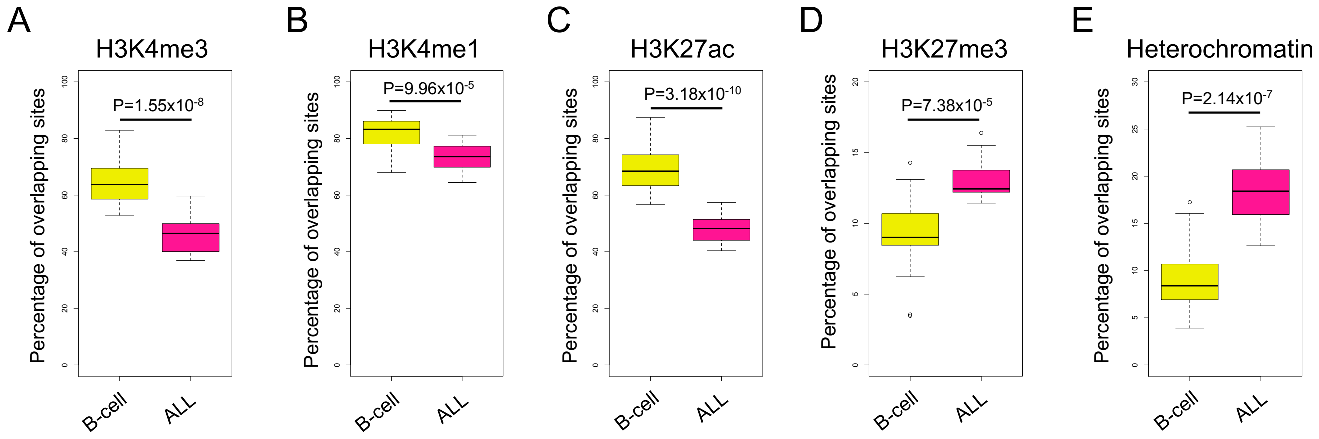
SNP genotyping data from patient ALL samples was obtained from published studies (17). ALL subtype caQTLs were identified by WASP (22) using ATAC-seq and SNP genotyping data from 24 primary ALL cell samples (n=6 for ETV6-RUNX1-specific caQTL analyses, n=7 for DUX4/ERG-specific caQTL analyses and n=11 for Hyperdiploid-specific caQTL analyses). We ran WASP on 932,868 genotyped SNPs and 9,539,719 imputed SNPs (R^2>0.6, MAF>1%; using Michigan Imputation Server). We further ensured that all genotypes SNPs were in Hardy-Weinberg Equilibrium, and we further only used polymorphic (i.e. heterozygous) SNPs in each patient biospecimen (6,627,136 polymorphic SNPs on average identified in each subtype). Somatic variants found in ETV6-RUNX1, DUX4/ERG and hyperdiploid subtypes were obtained from the Pediatric Cancer Genome Project (PCGP) (23). Pathogenicity scores were determined using Combined Annotation Dependent Depletion (CADD) (24). A general linear model was used to determine statistical significance in pathogenicity scores between groups of somatic variants.

REFERENCES

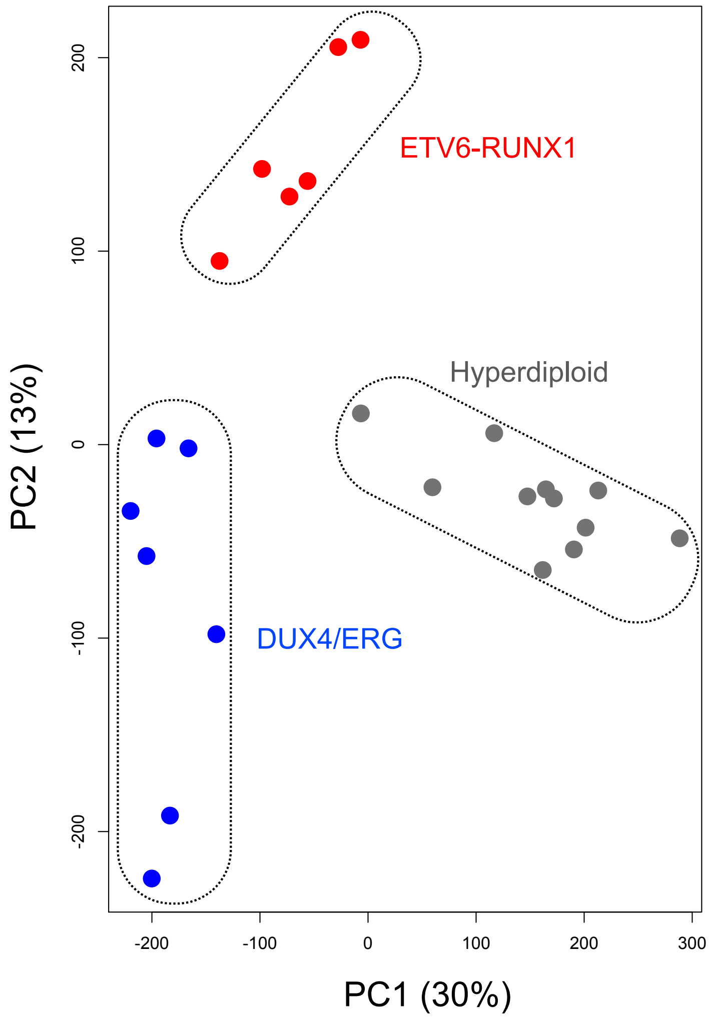
1. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016;48(10):1193-203.
2. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.
3. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
4. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
5. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
6. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome Biol. 2019;20(1):45.

7. Zhang J, McCastlain K, Yoshihara H, Xu B, Chang Y, Churchman ML, et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nat Genet. 2016;48(12):1481-9.
8. Tian L, Shao Y, Nance S, Dang J, Xu B, Ma X, et al. Long-read sequencing unveils IGH-DUX4 translocation into the silenced IGH allele in B-cell acute lymphoblastic leukemia. Nat Commun. 2019;10(1):2789.
9. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, 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):15545-50.
10. Savic D, Ramaker RC, Roberts BS, Dean EC, Burwell TC, Meadows SK, et al. Distinct gene regulatory programs define the inhibitory effects of liver X receptors and PPARG on cancer cell proliferation. Genome Med. 2016;8(1):74.
11. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
12. Calderon D, Nguyen MLT, Mezger A, Kathiria A, Muller F, Nguyen V, et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat Genet. 2019;51(10):1494-505.
13. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology. 2010;28(8):817.
14. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43.
15. Nordlund J, Backlin CL, Wahlberg P, Busche S, Berglund EC, Eloranta ML, et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. Genome Biol. 2013;14(9):r105.
16. Nordlund J, Backlin CL, Zachariadis V, Cavelier L, Dahlberg J, Ofverholm I, et al. DNA methylation-based subtype prediction for pediatric acute lymphoblastic leukemia. Clin Epigenetics. 2015;7:11.
17. Paugh SW, Bonten EJ, Savic D, Ramsey LB, Thierfelder WE, Gurung P, et al. NALP3 inflammasome upregulation and CASP1 cleavage of the glucocorticoid receptor cause glucocorticoid resistance in leukemia cells. Nat Genet. 2015;47(6):607-14.
18. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proc Natl Acad Sci U S A. 2017;114(25):E4914-E23.
19. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504.
20. Sentmanat MF, Peters ST, Florian CP, Connelly JP, Pruett-Miller SM. A Survey of Validation Strategies for CRISPR-Cas9 Editing. Sci Rep. 2018;8(1):888.
21. Connelly JP, Pruett-Miller SM. CRIS.py: A Versatile and High-throughput Analysis Program for CRISPR-based Genome Editing. Sci Rep. 2019;9(1):4194.
22. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015;12(11):1061-3.
23. Downing JR, Wilson RK, Zhang J, Mardis ER, Pui CH, Ding L, et al. The Pediatric Cancer Genome Project. Nat Genet. 2012;44(6):619-22.
24. 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):310-5.

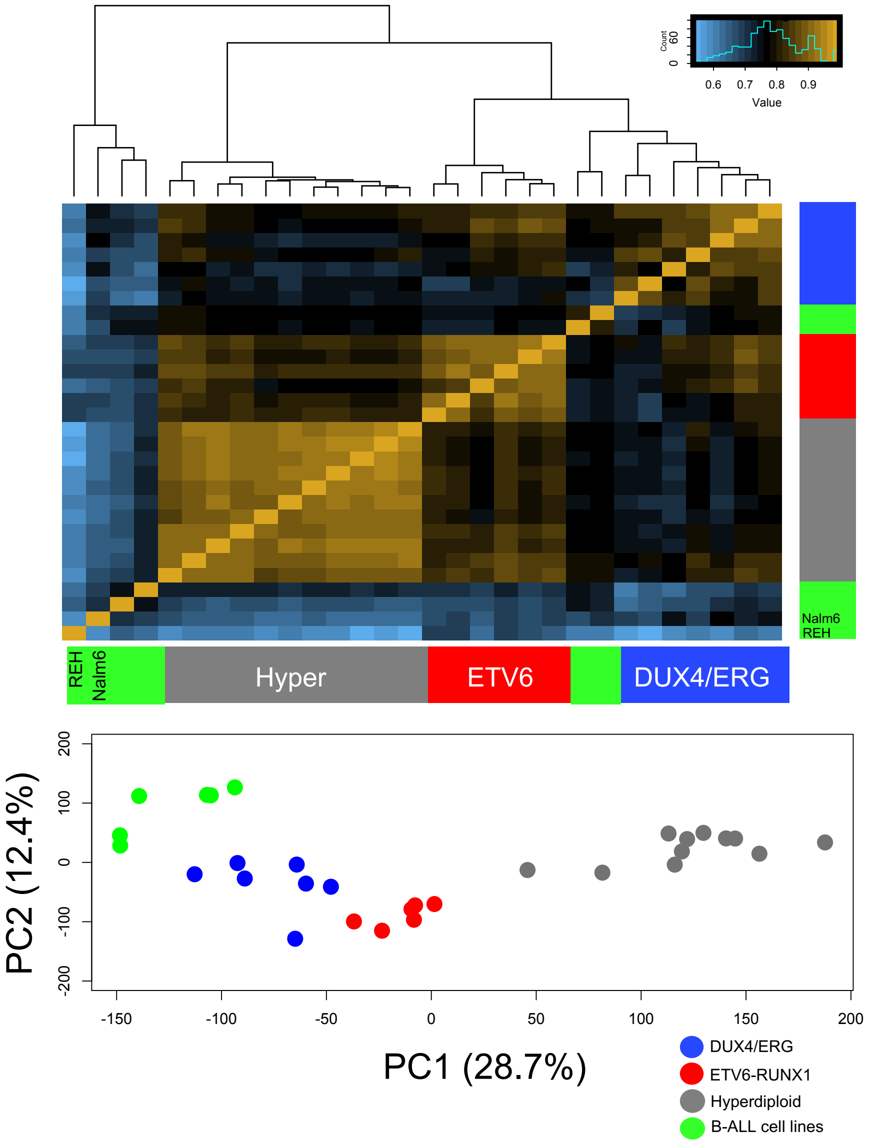
SUPPLEMENTAL FIGURES



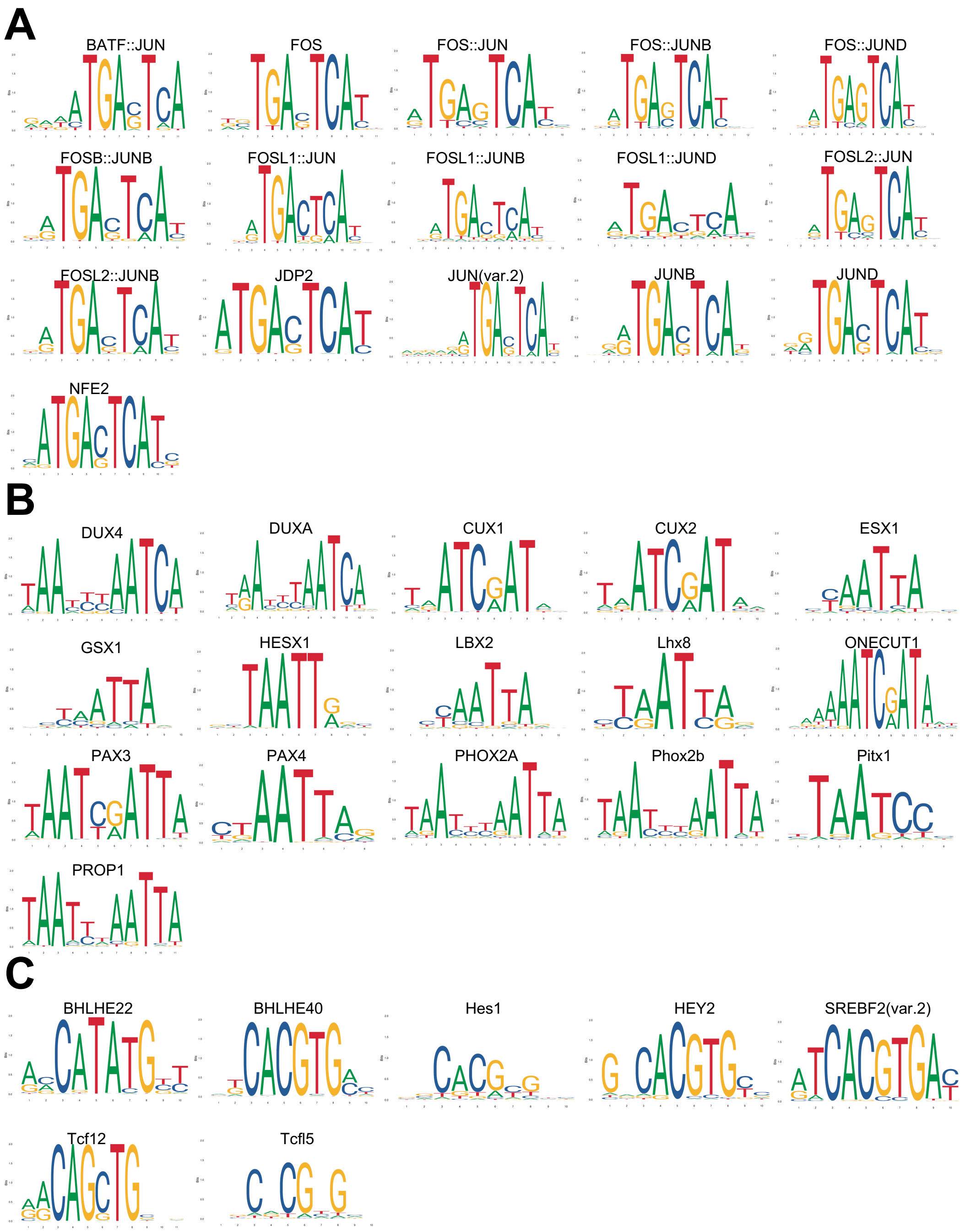
Supplemental Figure 1. Differences in chromatin state between ALL and normal B-cell open chromatin sites. Boxplots show differences in the percentage of overlapping open chromatin sites between ALL cells (n=24) and normal B-cells (n=15) with chromatin sites from normal B-cells from the Blueprint Epigenome consortium. Data for H3K4me3 (A), H3K4me1 (B), H3K27ac (C), H3K27me3 (D) and heterochromatin (E) is provided. Regions of heterochromatin were determined by identifying overlapping sites not mapping to H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3 sites in normal B-cells.



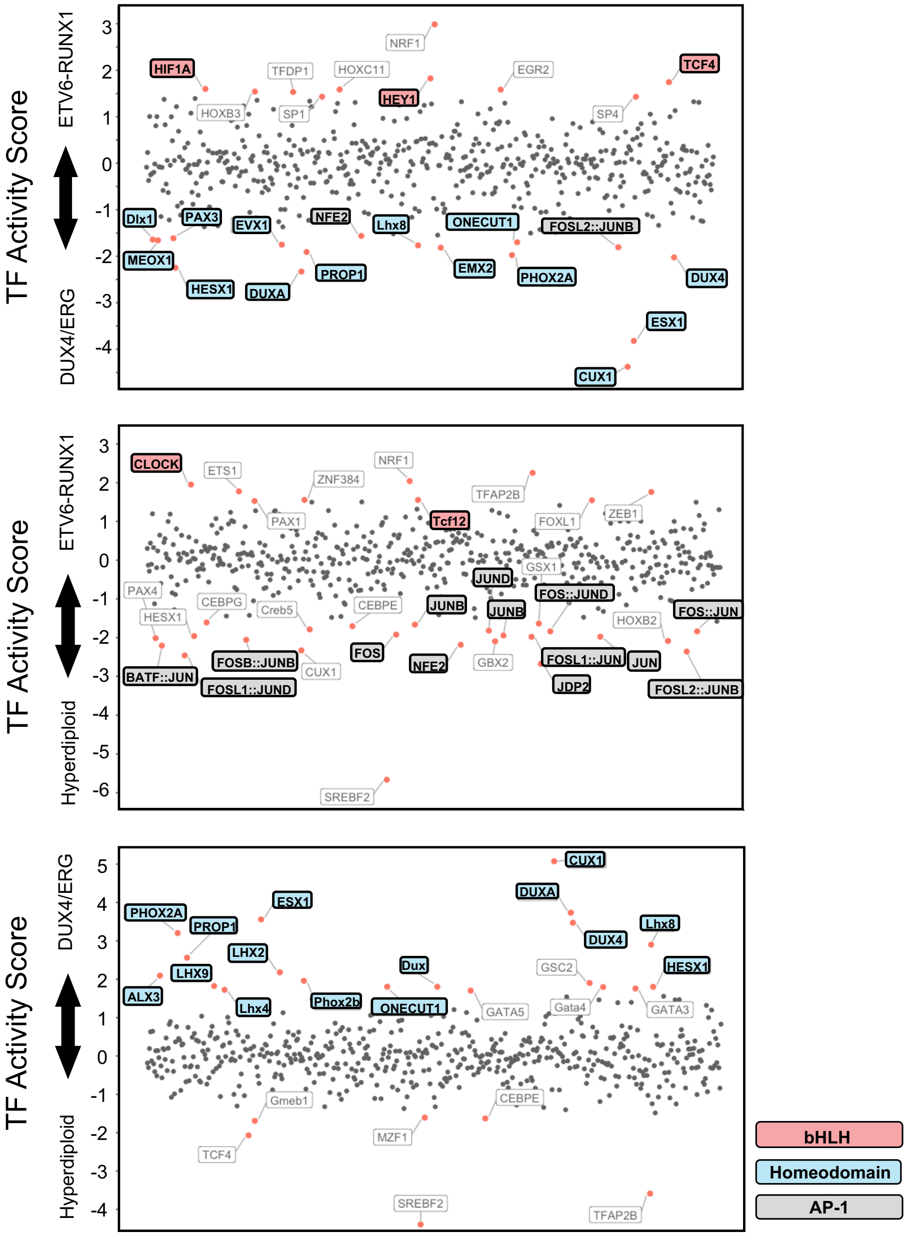
Supplemental Figure 2. PCA of all open chromatin sites. PCA plot of normalized read depth at all identified open chromatin sites in the ALL genome is shown.



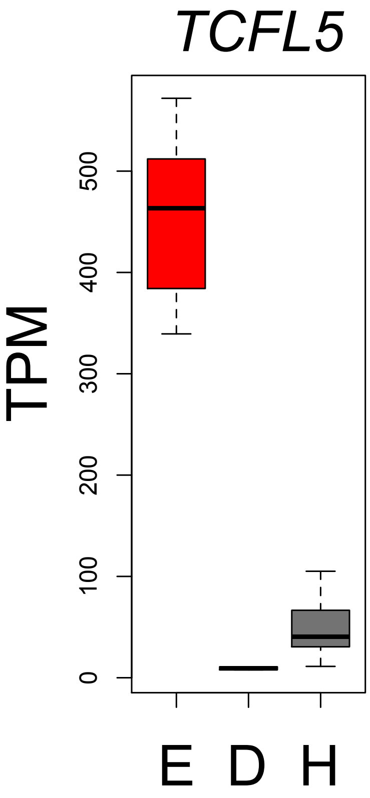
Supplemental Figure 3. Open chromatin landscapes of primary ALL cells and immortalized ALL cell lines. Above, unsupervised hierarchical clustering heatmap of primary ALL cells and immortalized ALL cell lines of normalized read depth at all overlapping open chromatin sites between primary ALL cell sand B-ALL cell lines (76267 sites) is depicted. ALL cell lines are depicted in green and locations of ETV6-RUNX1 REH cells and DUX4/ERG-like Nalm6 cells are provided. Below, PCA plot of normalized read depth at all overlapping open chromatin sites between primary ALL cell and B-ALL cell line sites (76267 sites) is shown below.



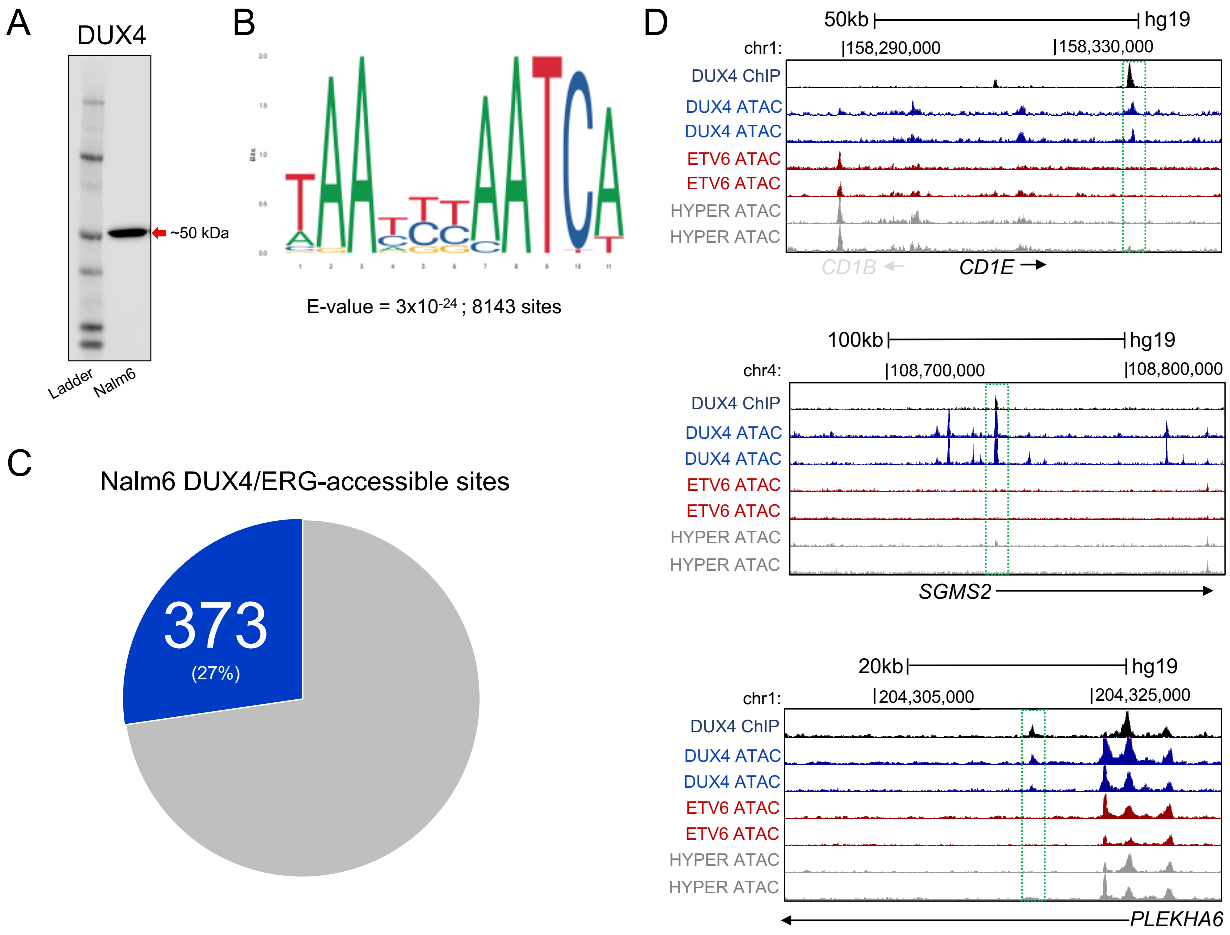
Supplemental Figure 4. Significant transcription factor activity score motif. JASPAR (http://jaspar.genereg.net/) motifs from TFs exhibiting significant TF activity scores from pairwise ALL subtype comparisons. Motifs for TF families that are consistently enriched in hyperdiploid (A), DUX4/ERG (B) and ETV6-RUNX1 (C) are shown.



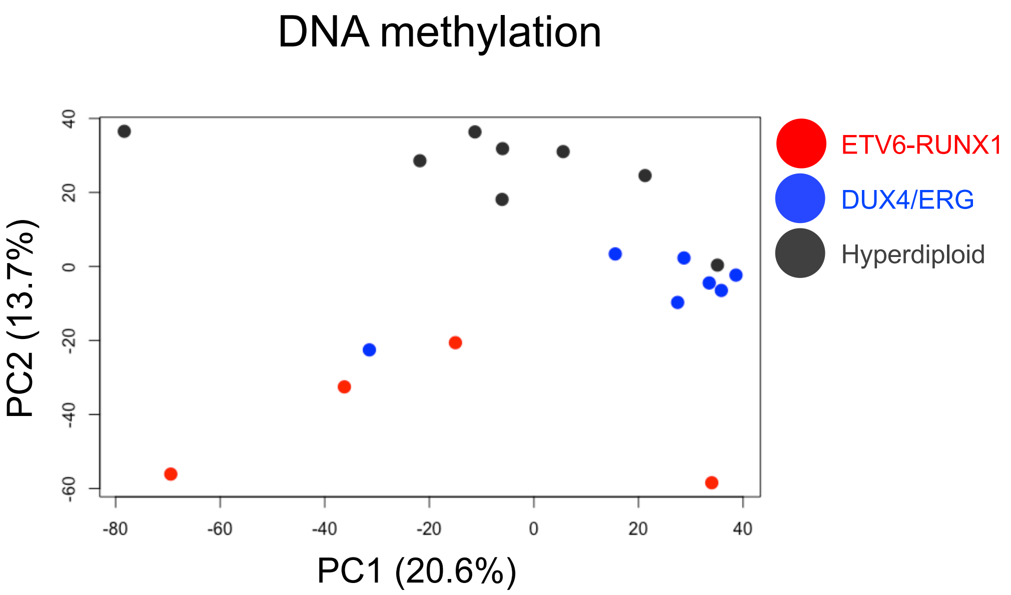
Supplemental Figure 5. TF activity score comparisons at ALL subtype-accessible sites (A) Dot plots show ALL subtype comparisons of TF activity scores (x-axis; $p<0.05$) at ALL subtype-accessible sites for all pairwise analyses. TF families and corresponding colors are given at the bottom right.



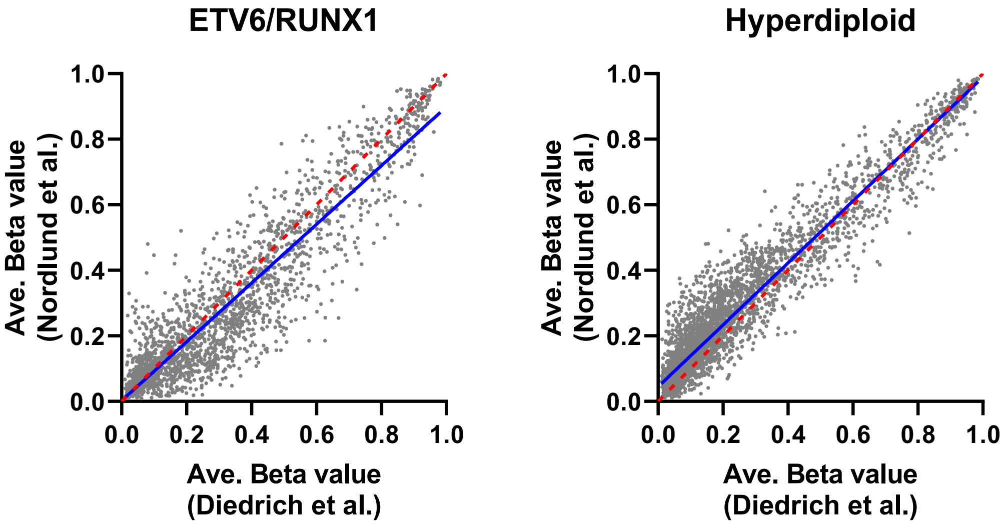
Supplemental Figure 6. Expression levels of candidate bHLH TF gene. Average transcripts per million (TPM) for *TCFL5* is shown among ALL subtypes (E=ETV6-RUNX1, D=DUX4/ERG, H=Hyperdiploid).



Supplemental Figure 7. DUX4 ChIP-seq in Nalm6 cells. (A) Western blot identifies DUX4 protein expression in Nalm6 cells. (B) DUX4 ChIP-seq was significantly enriched in canonical DUX4 motif (JASPAR MA0468.1, E-value = 3x10^{-24}, 8143 sites). (C) 373 of 1367 DUX4/ERG subtype-accessible sites that were preferentially accessible in Nalm6 cells (i.e. Nalm6 DUX4/ERG subtype-accessible sites; see Fig. 2D) were bound by DUX4. 749 total DUX4/ERG subtype-accessible chromatin sites with evidence of DUX4 binding in Nalm6 cells. (D) Examples of DUX4/ERG-accessible sites bound by DUX4 (outlined in green).

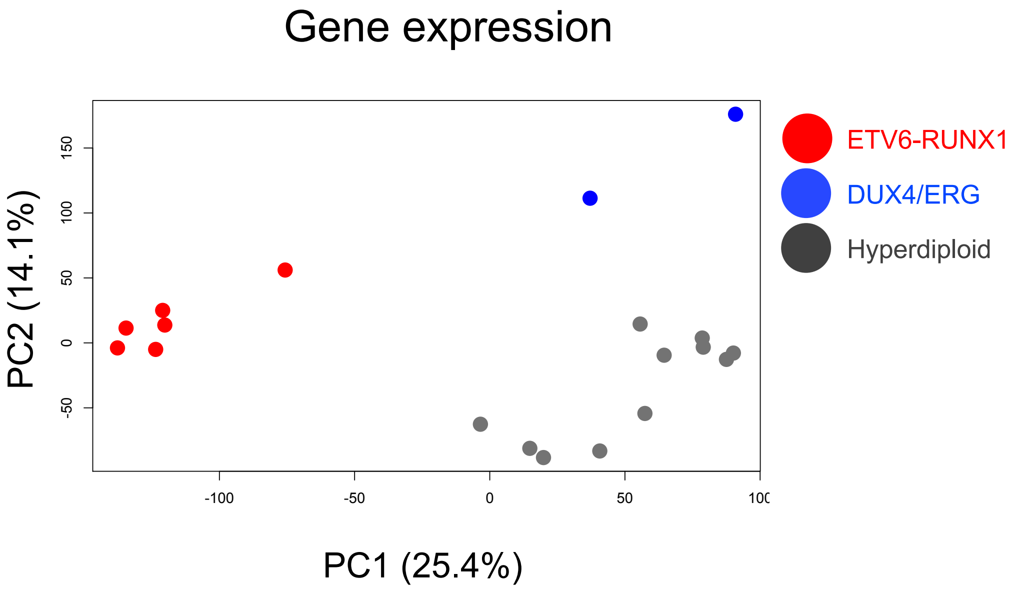


Supplemental Figure 8. Principal component analysis of DNA methylation. Principal component analysis using CpG DNA methylation beta-values on DNA methylation array.

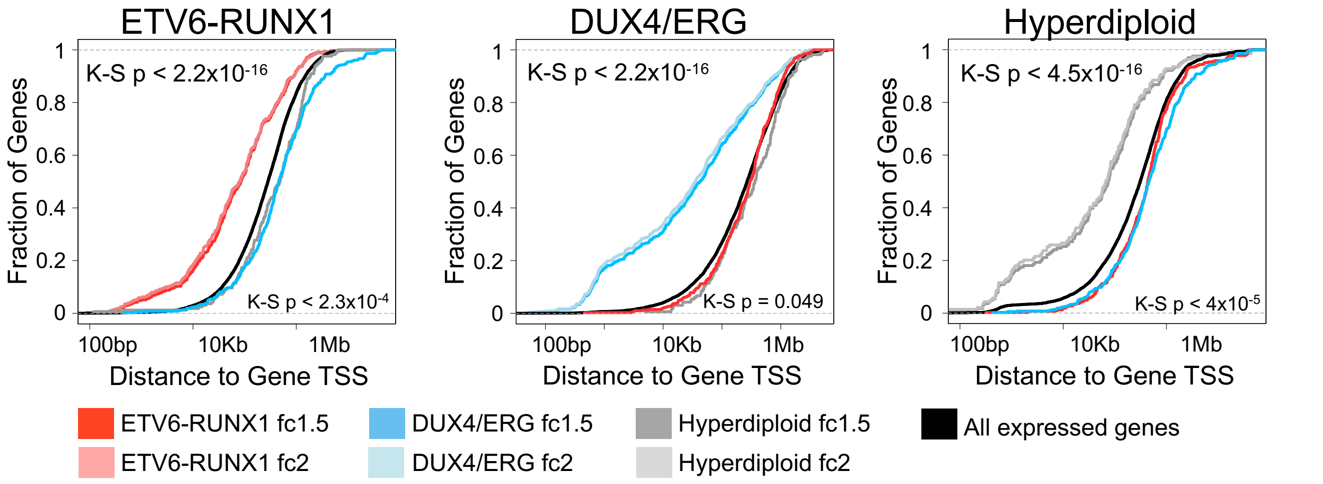


Subtype	GpG sites compared	Sample size (Diedrich et al.)	Sample size (Nordlund et al.)	Spearman rho (p-value)
ETV6-RUNX1	2474	4	30	0.91 (<0.0001)
Hyperdiploid	4116	8	30	0.94 (<0.001)

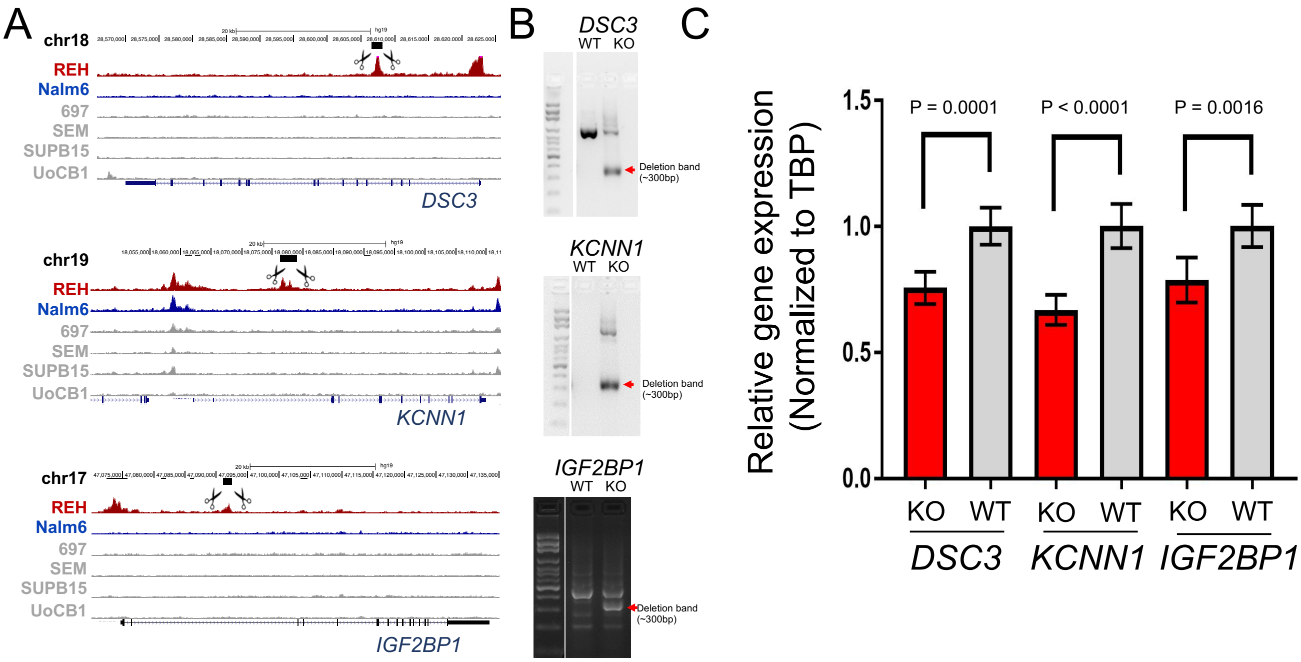
Supplemental Figure 9. Correlation of CpG DNA methylation at subtype-accessible chromatin sites. Plots depict the correlation between average CpG probe beta values from Diedrich *et al.* and 30 randomly selected ALL biospecimens from an independent cohort (Nordlund *et al.,* GEO ID: GSE49031) for two distinct ALL subtypes, ETV6/RUNX1 (left) and Hyperdiploid (right). In each plot the grey data points represent individual CpG probes, the red dashed line shows perfect correlation, and the blue line shows the fitted linear correlation. Below, a table displays the total number of datasets used, and provides Spearman *rho* correlations and significance.



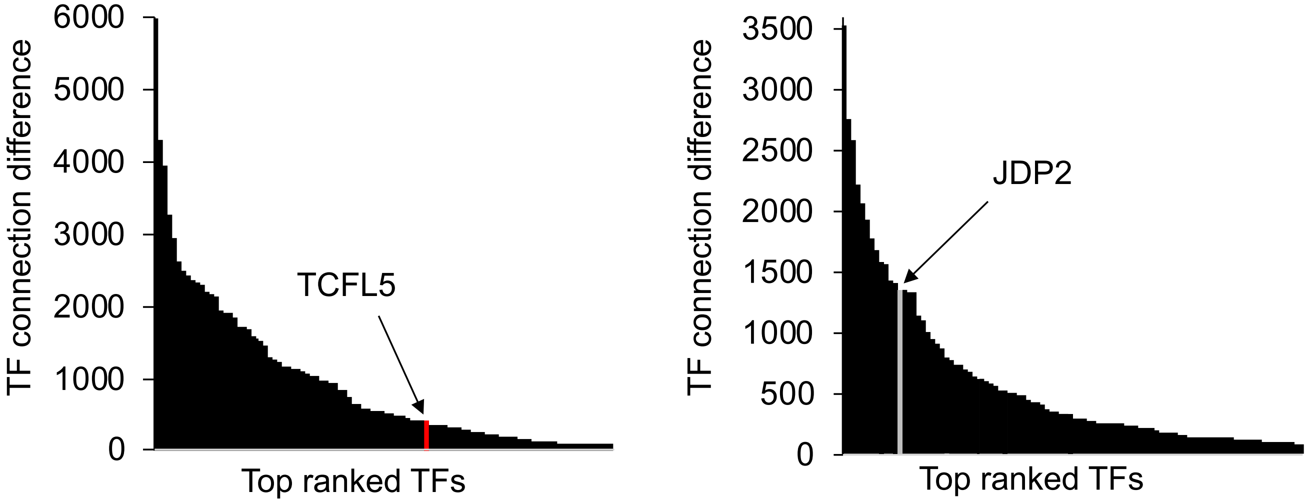
Supplemental Figure 10. Principal component analysis of gene expression. Principal component analysis using normalized gene counts from RNA-seq experiments.



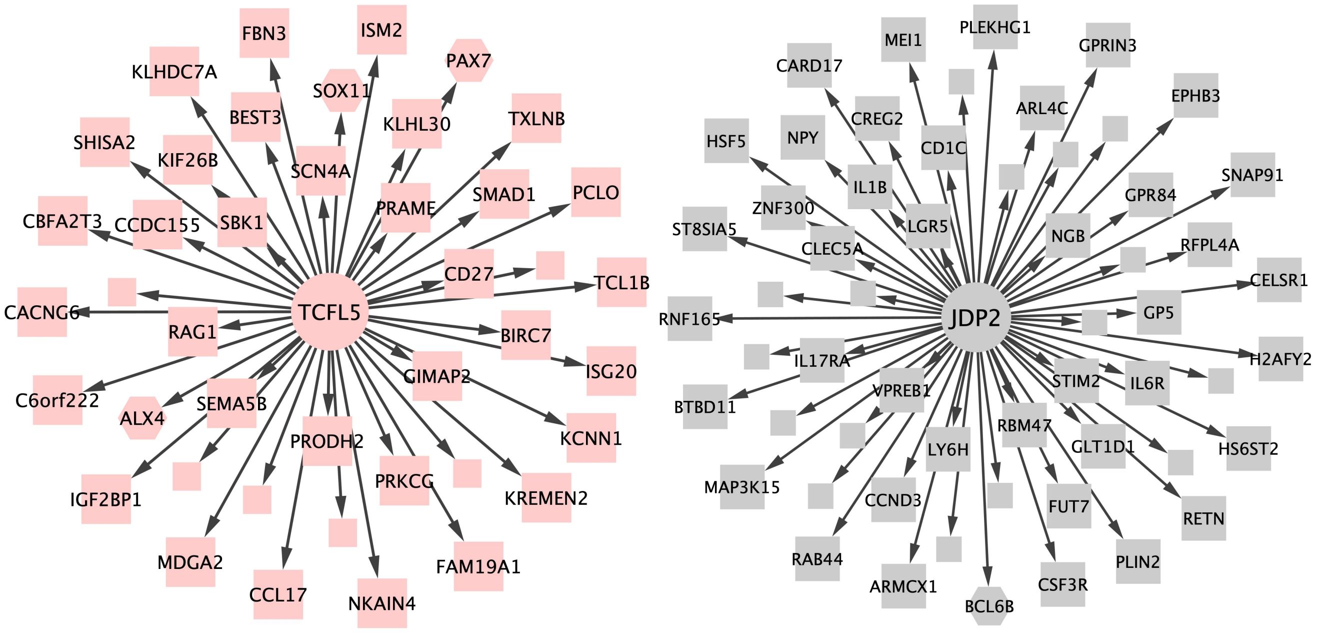
Supplemental Figure 11. Enrichment of subtype-depleted sites near down-regulated DEGs. Cumulative distribution functions display the fraction (y-axis) and distance (x-axis) of ALL subtype down-regulated gene transcription start sites to the nearest subtype-depleted site. Data is provided for ETV6-RUNX1-depleted sites, DUX4/ERG-depleted sites and hyperdiploid-depleted sites (left-to-right). For each set of subtype-depleted sites, distances to genes down-regulated >2-fold (fc2) or >1.5-fold (fc1.5) within the same subtype are provided, as well as genes down-regulated >1.5-fold in opposing subtypes. Background distance distributions use transcription start sites of all expressed genes in each subtype. Kolmogorov-Smirnov (K-S) test p-values are provided for each plot. In the upper left are K-S p-values showing enrichment for the same subtype, while the bottom right are K-S p-values showing deletion for opposing subtypes. For DUX4/ERG, only hyperdiploid down-regulated genes showed a significant depletion (p = 0.049) with DUX4/ERG-depleted sites.



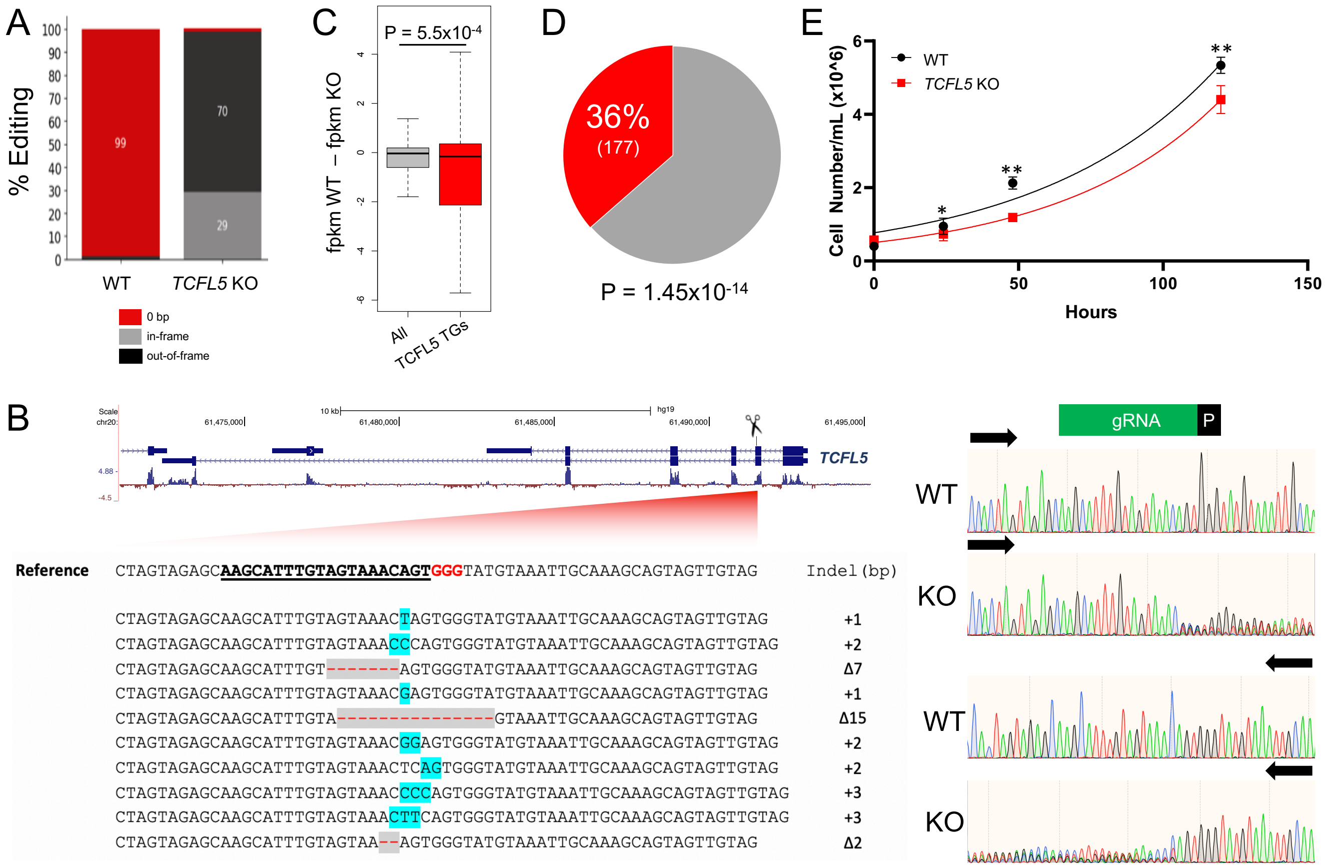
Supplemental Figure 12. CRISPR/Cas9 KO of ETV6-RUNX1 subtype-accessible sites. (A) Images of *DSC3*, *KCNN1* and *IGF2BP1* gene loci. ETV6-RUNX1 subtype-accessible sites are shown (black boxes flanked by scissors). (B) PCR validation of ETV6-RUNX1 subtype-accessible site deletion using PCR validation. (C) RT-qPCR from RUNX1 subtype-accessible sites knockout REH (KO) and parental REH (WT) cells are shown. P-values are given.



Supplemental Figure 13. TF connection ranking in ALL subtypes. Distribution of top TFs ranked by enrichment in target gene connections in ETV6-RUNX1 ALL (left) and hyperdiploid ALL (right). Locations of TCFL5 (in red) and JDP2 (in gray) are denoted in the TF ranking.

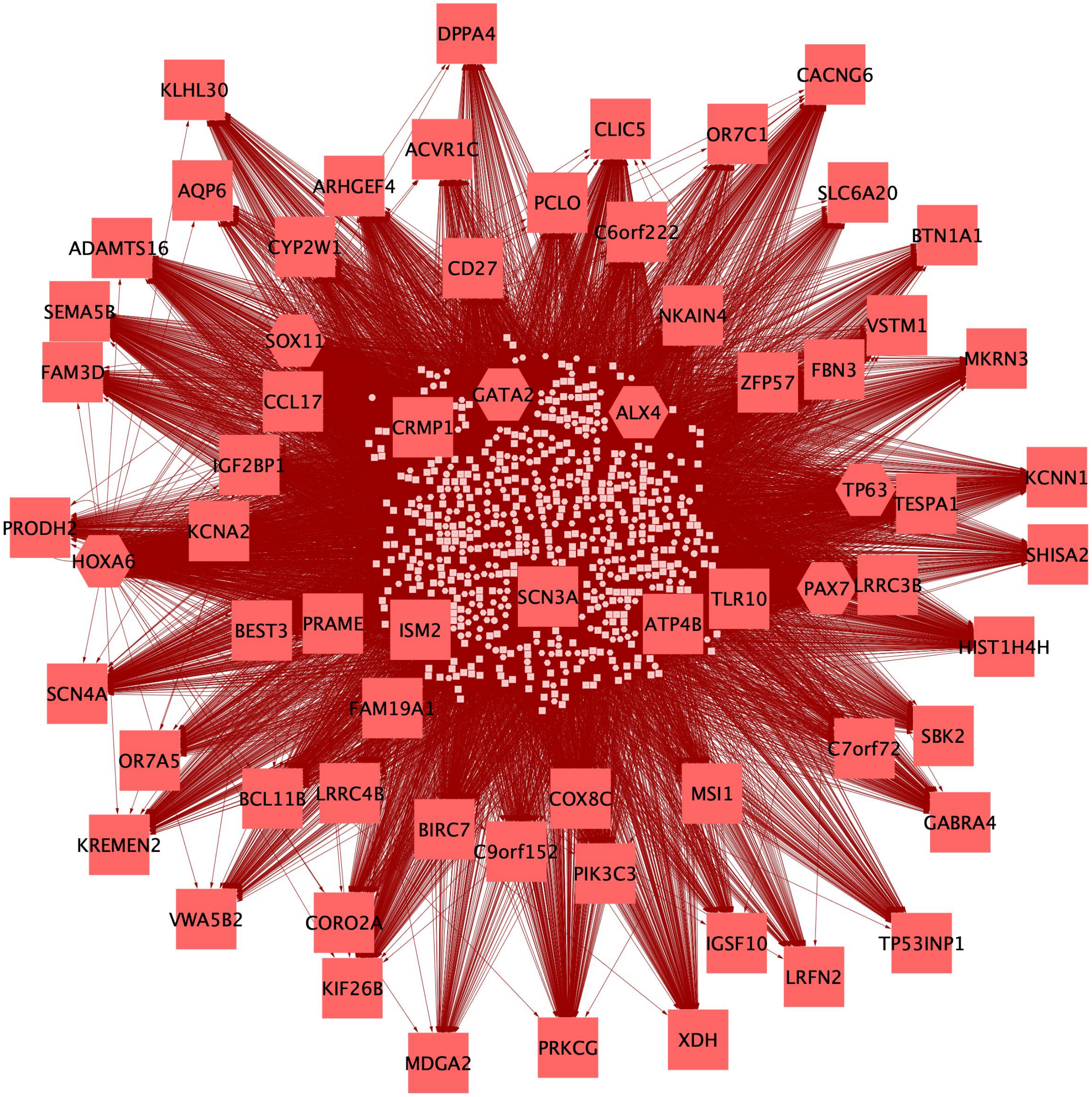


Supplemental Figure 14. TCFL5 and JDP2 upregulated DEG target genes. Gene regulatory network map of TCFL5 (left) and JDP2 (right) TF connections to target genes that are subtype-specific up-regulated DEGs in ETV6-RUNX1 ALL and hyperdiploid ALL respectively. Target gene names are provided for DEGs with an average log2 fold change greater than 2. Target genes that also act as TFs are shown as hexagons.

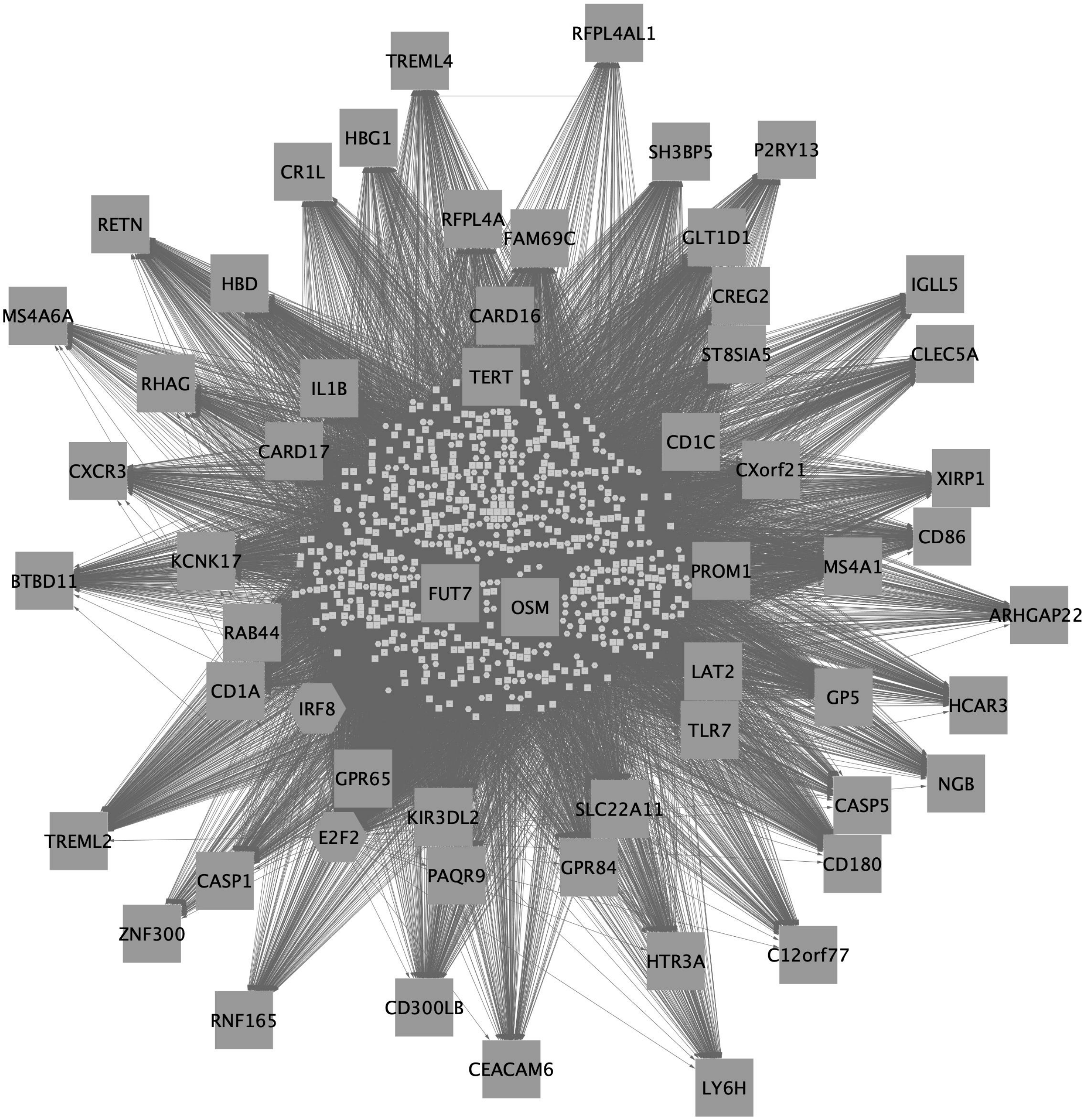


		Indel (bp)
Reference	CTAGTAGAGC**AAGCATTTGTAGTAAACAGT**GGGTATGTAAATTGCAAAGCAGTAGTTGTAG	
	CTAGTAGAGCAAGCATTTGTAGTAAACTAGTGGGTATGTAAATTGCAAAGCAGTAGTTGTAG	+1
	CTAGTAGAGCAAGCATTTGTAGTAAACCCAGTGGGTATGTAAATTGCAAAGCAGTAGTTGTAG	+2
	CTAGTAGAGCAAGCATTTGT-------AGTGGGTATGTAAATTGCAAAGCAGTAGTTGTAG	Δ7
	CTAGTAGAGCAAGCATTTGTAGTAAACGAGTGGGTATGTAAATTGCAAAGCAGTAGTTGTAG	+1
	CTAGTAGAGCAAGCATTTGTA---------------GTAAATTGCAAAGCAGTAGTTGTAG	Δ15
	CTAGTAGAGCAAGCATTTGTAGTAAACGGAGTGGGTATGTAAATTGCAAAGCAGTAGTTGTAG	+2
	CTAGTAGAGCAAGCATTTGTAGTAAACTCAGTGGGTATGTAAATTGCAAAGCAGTAGTTGTAG	+2
	CTAGTAGAGCAAGCATTTGTAGTAAACCCCAGTGGGTATGTAAATTGCAAAGCAGTAGTTGTAG	+3
	CTAGTAGAGCAAGCATTTGTAGTAAACTTCAGTGGGTATGTAAATTGCAAAGCAGTAGTTGTAG	+3
	CTAGTAGAGCAAGCATTTGTAGTAA--AGTGGGTATGTAAATTGCAAAGCAGTAGTTGTAG	Δ2

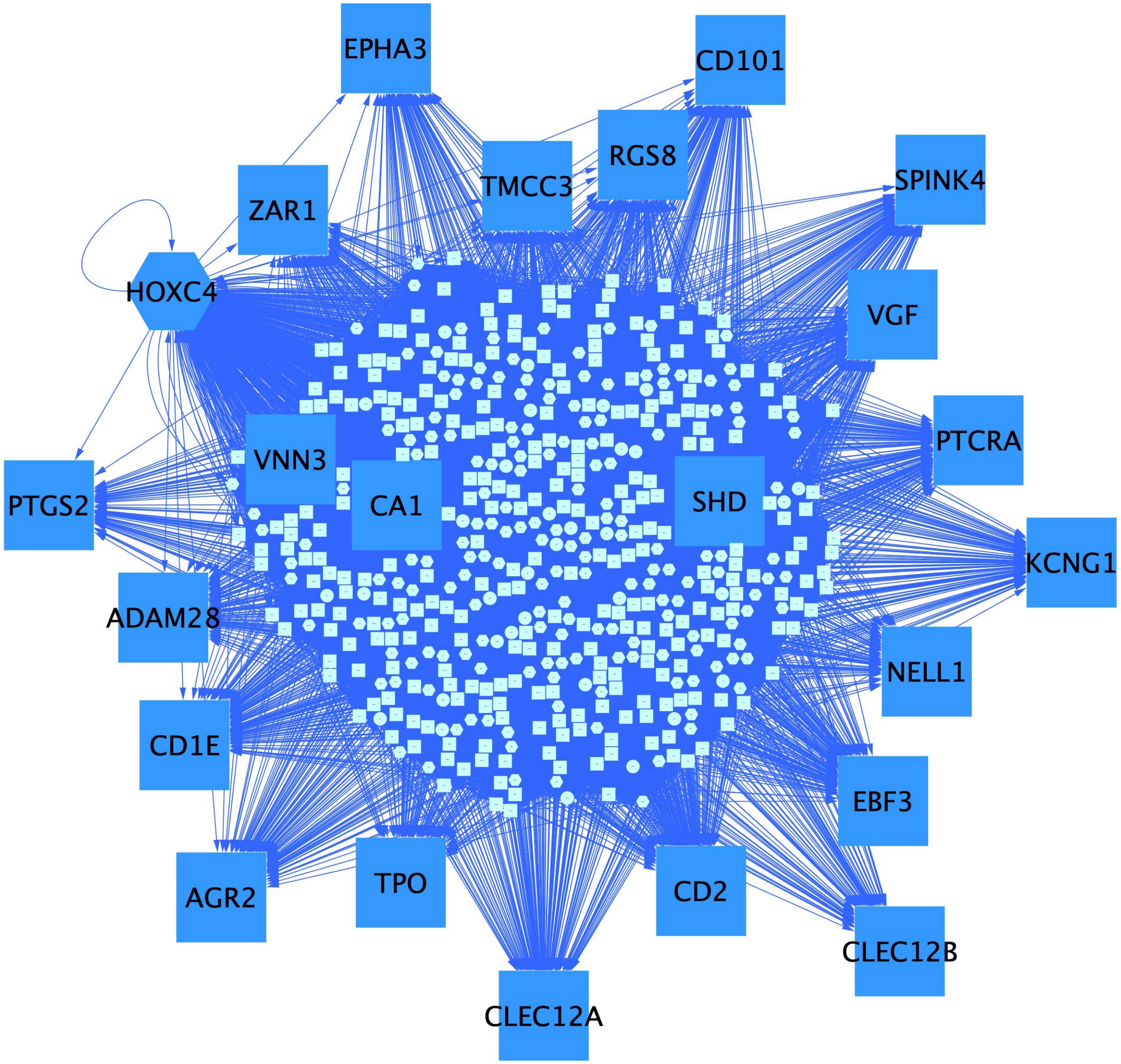
Supplemental Figure 15. CRISPR/Cas9 disruption of *TCFL5* in REH cells. CRISPR/Cas9 genome editing was used to disrupt *TCFL5.* (A) Next-generation sequencing (NGS) was used to determine the percentage of wild-type (WT, 0-bp; red) in-frame mutations (gray) and out-of-frame mutations (black) in WT parental REH cells and *TCFL5* disrupted (*TCFL5* KO) cells. (B) Next-generation sequence alignment of mutant alleles from hTCFL5 knockdown pool (left). The *TCFL5* gene locus and location of disruption is highlighted. Below, the sgRNA target sequence is underlined and bolded and the PAM sequence is in red. Deletions are represented by dashes and insertions are highlighted in turquoise. The size of each indel is shown on the right. Sanger sequencing results in WT and *TCFL5* KO cells are provided in the right panel. (C) Boxplot showing a significant difference in gene expression (fpkm) between parental and *TCFL5* KO cells for expressed genes and predicted TCFL5 target genes from gene regulatory network analyses is shown. (D) A significant fraction of differentially expressed genes between parental and TCFL5 KO REH cells (FDR<0.01) are predicted TCFL5 target genes (in red; 36%; Fisher's Exact Test p=1.45x10-14; 2.2-fold enriched). (E) Cellular proliferation assays in WT and TCLF5 KO cells show significant decreases in proliferation after *TCFL5* disruption. Cell count data was normalized to the 0hr timepoint to test for significance (*p<0.05, **p<0.01).



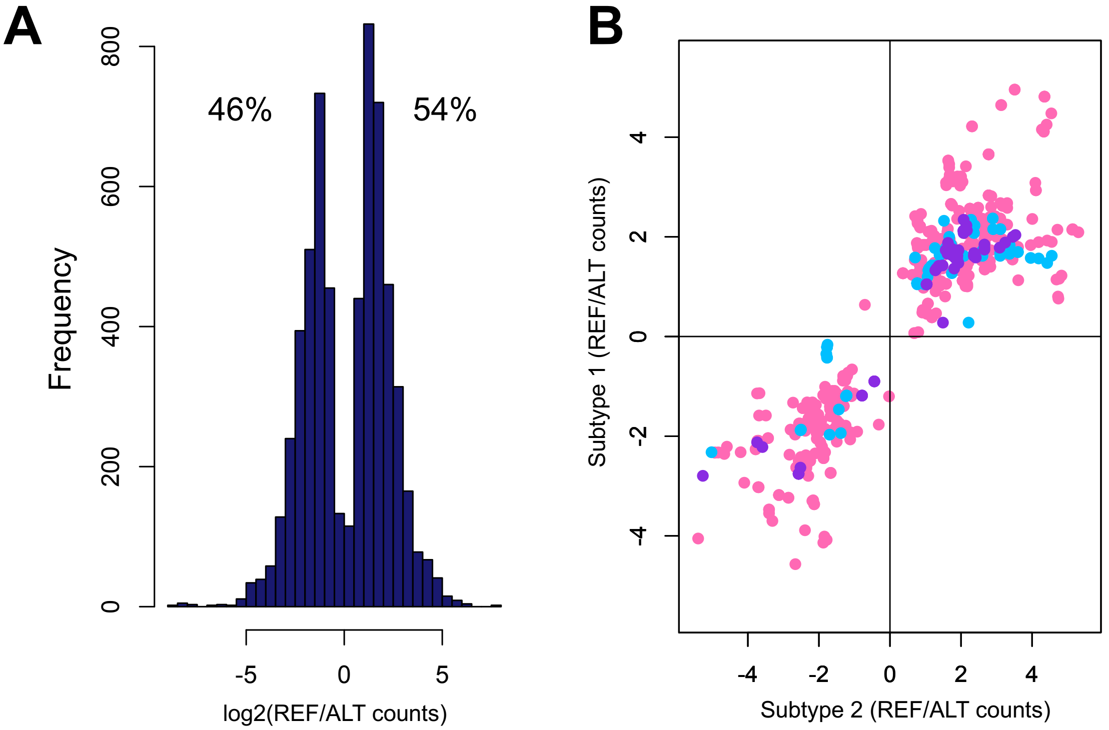
Supplemental Figure 16. ETV6-RUNX1 gene regulatory network of target genes. Gene regulatory network map of enriched ETV6-RUNX1 ALL target genes is provided. Gene names highlight enriched target genes that reproducibly exhibited >=75 connections in ETV6-RUNX1 ALL compared to the two opposing subtypes. The network was further pruned to only show network hubs with >150 total network interactions that were directly connected to these enriched target genes. TF interactors are shown as circles, target genes are presented as squares and TFs identified as both TF interactors and target gene interactors are shown as hexagons. Arrows stem from TFs and point to gene targets.



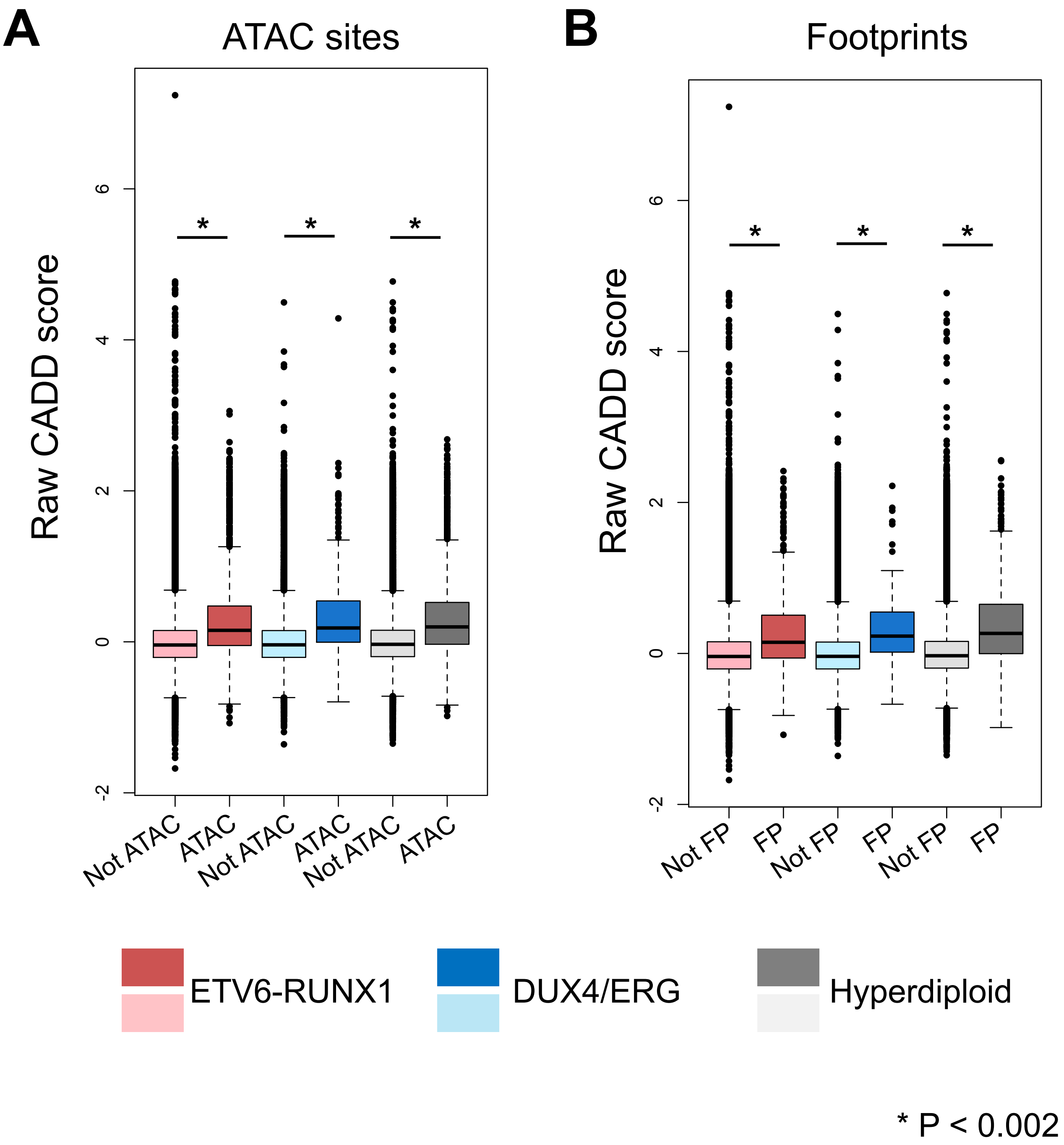
Supplemental Figure 17. Hyperdiploid gene regulatory network of target genes. Gene regulatory network map of enriched hyperdiploid ALL target genes is provided. Gene names highlight enriched target genes that reproducibly exhibited >=75 connections in hyperdiploid ALL compared to the two opposing subtypes. The network was further pruned to only show network hubs with >150 total network interactions that were directly connected to these enriched target genes. TF interactors are shown as circles, target genes are presented as squares and TFs identified as both TF interactors and target gene interactors are shown as hexagons. Arrows stem from TFs and point to gene targets.



Supplemental Figure 18. DUX4/ERG gene regulatory network of target genes. Gene regulatory network map of enriched DUX4/ERG ALL target genes is provided. Gene names highlight enriched target genes that reproducibly exhibited >=75 connections in DUX4/ERG ALL compared to the two opposing subtypes. The network was further pruned to only show network hubs with >150 total network interactions that were directly connected to these enriched target genes. TF interactors are shown as circles, target genes are presented as squares and TFs identified as both TF interactors and target gene interactors are shown as hexagons. Arrows stem from TFs and point to gene targets.



Supplemental Figure 19. Chromatin accessibility QTL analyses (FDR<0.1). (A) Log distribution of allele-specific accessibility ratios (reference versus alternative allele count) at caQTLs (FDR<0.1) in ALL cell samples (REF=reference, ALT=alternative). **(B)** Comparison of log2 allele-specific accessibility read count ratios (REF/ALT) between overlapping ALL subtype caQTLs (FDR<0.1). Hyperdiploid (X-axis) and ETV6-RUNX1 (Y-axis) overlapping caQTLs are shown in pink, hyperdiploid (X-axis) and DUX4/ERG (Y-axis) overlapping caQTLs are shown in light blue and ETV6-RUNX1 (X-axis) and DUX4/ERG (Y-axis) overlapping caQTLs are shown in purple.



Supplemental Figure 20. CADD score comparisons at accessible sites and TF footprints. Box plots show raw CADD scores for somatic variants in closed versus accessible chromatin in each ALL subtype (A), and for somatic variants in TF footprints versus outside of TF footprints in each ALL subtype (B). Not ATAC = closed chromatin; ATAC = accessible chromatin; Not FP = not in TF footprint; FP = in TF footprint. Significant differences (marked by asterisk; p<0.002) are shown for all subtypes.