SUPPLEMENTARY MATERIALS

Comprehensive genomic analysis reveals FLT3 activation and a therapeutic strategy for a patient with relapsed adult B lymphoblastic leukemia

Short running title: Sequencing reveals *FLT3* activation in an adult B-ALL case

Inventory of Supplementary Materials:

Supplementary Methods
Extended descriptions of experimental and analysis methods.

Supplementary Results
Extended descriptions of genomic and transcriptomic analysis results.

Supplementary Tables
Supplementary Tables S1-S8. All supplementary tables are available as spreadsheets that can be downloaded from the journal website.

All sequence data described in this publication including the whole genome, exome, RNA-seq and custom capture data have been deposited in the Database of Genotypes and Phenotypes (dbGAP) (accession: phs001066.v1.p1; **Table S8** for details).

Supplementary Figures
Supplementary Figures S1-S41.

Author Contributions

References

SUPPLEMENTARY METHODS

Sample acquisition, nucleic acid isolation and patient information

Results in this work correspond to tissue samples obtained from the patient at nine distinct time points spanning 4,024 days from diagnosis (see **Table S1** for details and **Figure 1** for an overview of the timeline). All samples were annotated with a time point value representing the number of days from diagnosis (day 0). A skin sample was obtained as a normal reference at 42 days from diagnosis. At this day, the patient was considered to be in remission with no detectable tumor burden. Samples representing the primary tumor (day 0) were obtained from a clot obtained from the bone marrow aspirate (annotated as "day 0, clot") and from a fixed slide of bone marrow cells that were obtained from the Wright-Giemsa stained bone marrow aspirate slide (annotated as "day 0, slide"). Both of these samples were of marginal quantity and quality. All additional tissue samples described in this work were obtained from total bone marrow aspirate or core (the solid core left after removal of the aspirate). Marrow core samples were decalcified and then fixed and placed in paraffin for sectioning prior to further analysis. The aspirates were never fixed and were processed to cell pellets and frozen prior to additional analysis. Samples were obtained during first remission (day 42), first relapse (day 1,893), second relapse (days 3,068 and 3,072), and during final remission (day 3,219 and 4,024). The patient received a bone marrow allograft from a sibling (brother) on day 2,010 after achieving a second remission. He received a second bone marrow allograft from a matched unrelated donor (MUD) on day 3,151 after achieving a third remission. Since the second relapse sample occurred after the first allograft from the sibling donor we subjected this sample to sorting for blasts to enrich for tumor cells from the patient. As a control we separately sorted for lymph cells to enrich for cells from the sibling donor. Blasts and lymphocytes were sorted concurrently from the same second relapse sample by CD45 and side scatter (lymphocytes were CD45 bright with low side scatter and blasts were CD45 dim with low side scatter) (see **Figure S2**). We then collected CD19+/CD34+ cells in the blast gate. Antibodies used were PerCP-CY5.5 CD45 (eBioscience; clone 2D1), PE-CD34 (PE-pool; Beckman Coulter Genomics; PN IM1459U) and FITC-CD19 (BD Biosciences; clone HIB19). The majority of variant discovery analysis was performed on the sorted blasts from the second relapse sample with validation/verification of these events being conducted in the additional time point samples. Blood from our B Lymphoblastic Leukemia (B-ALL) patient and both the sibling bone marrow donor and matched unrelated donor were subjected to standard HLA typing. The B-ALL patient had HLA type as follows: 'HLA-A31:01 / HLA-A02:01', 'HLA-B18:01 / HLA-B08:01', 'HLA-C07:01'. Both the sibling donor and unrelated donor had matching HLA type. All of the samples described above were subjected to genomic DNA isolation by column purification (Qiagen DNeasy). The sorted blasts from day 3,072 sample was also subjected to RNA isolation using a TRIzol® procedure followed by DNAseI treatment and cleanup using a Qiagen RNeasy mini kit.

Cytogenetics and quantitative interphase FISH

Cytogenetics analyses were performed on marrow samples obtained for the primary diagnosis (day 0), first relapse (day 1,893), second relapse (day 3,072), and additional time points during remission. FISH was performed using probes for *KMT2A* (also known as *MLL*), *ETV6/RUNX1* (also known as *TEL/AML*), and *BCR/ABL1*. Standard cytogenetics analysis using GTG banding was performed on each disease sample.

Whole genome, exome, and transcriptome sequencing

Whole genome, exome and transcriptome sequencing (RNA-seq) were performed using library construction and capture hybridization methods previously described [1, 2]. All sequence data were generated as 2x100 bp reads on the Illumina HiSeq 2000 or HiSeq 2500 platforms. Whole genome sequence libraries were constructed from two DNA fragment size ranges each, for both tumor (~150-250 bp; ~300-450 bp) and normal (~100-300 bp; ~100-450 bp). Exome sequence libraries were constructed independently from the whole genome libraries from DNA fragments size selected to be ~100-400 bp. Exome capture hybridizations used the NimbleGen SeqCap EZ Human Exome Library v2.0 (for the skin normal and second relapse) or NimbleGen

SeqCap EZ Human Exome Library v3.0 (for the primary tumor and first relapse) kits (Roche, Inc.). Unstranded (non strand specific) RNA-seq libraries were constructed using polyA+ selected RNA input and the Ovation RNA-Seq System V2 (Nugen, Inc.) kit. cDNA fragmentation was performed using a Covaris instrument set to 'Broad Range Program 2' (5DC/4I/200CB/90sec). A size selection targeting fragments in the range of 300-500 bp was performed using a dual SPRI (Solid Phase Reversible Immobilisation) strategy. SPRI beads were obtained from AMPure.

DNA sequence alignment and data quality assessment

Alignment of whole genome and exome data was performed using the Genome Modeling System (GMS) essentially as described in Griffith et al. 2015 using the GMS processing profile: 'January 2015 Reference Alignment Candidate 1' [2]. Paired end read sequences were aligned to the human genome reference sequence (version GRCh37 also known as hg19). Read alignments were performed using bwa-mem [3] version 0.7.10 with default parameters except '-t 8' to utilize 8 threads for parallel processing. For datasets with multiple lanes of data, these were aligned independently for each lane and merged using Picard 'MergeSamFiles' version 1.113 with default parameters. Duplicate reads were marked with Picard 'MarkDuplicates' version 1.113 with default parameters. For datasets generated from multiple sequence libraries duplicates were marked within each library prior to merging into a final BAM file (e.g. WGS datasets each consisted of two independent fragment size libraries). Quality of the alignments was assessed by metrics determined by Samtools flagstat (version 0.1.19), FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), Picard 'CollectInsertSizeMetrics' (version 1.113), and Picard 'CollectWgsMetrics' (version 1.113). To assess sample quality and identify possible sample swaps, each sample was subjected to analysis on an Illumina iScan Instrument utilizing a Human OmniExpress genotyping array according to the manufacturer's recommendations (Illumina Inc, San Diego, CA). Genotypes obtained from this platform were assessed for concordance with SNP genotypes obtained by whole genome sequencing. Sample identities were further confirmed by WGS genotyping of 24 biallelic SNP locations previously selected for this purpose [4]. To obtain variant allele frequencies for these SNP positions, read counts supporting reference and variant alleles were obtained using bam-readcount v0.6 (https://github.com/genome/bam-readcount) and alternative allelic frequencies were compared using the GenVisR package in R (https://github.com/griffithlab/GenVisR).

Genome-wide sequencing depths in the WGS data were calculated using Picard CollectWgsMetrics (https://broadinstitute.github.io/picard/) with a minimum mapping (MINIMUM_MAPPING_QUALITY) and base quality (MINIMUM_BASE_QUALITY) equal to 20, a maximum coverage (COVERAGE_CAP) of 100,000, and lenient validation stringency of files (VALIDATION_STRINGENCY). Sequencing depths in targeted regions of the genome were calculated using SAMtools depth [5] with a minimum mapping (-q) and base quality (-Q) equal to 20. Cumulative coverage heatmaps were plotted using 'ggplot2' (http://ggplot2.org/) in conjunction with the 'GenVisR' package (https://github.com/griffithlab/GenVisR).

Germline variant analysis

Germline variant analysis was performed using WGS and exome data for the normal skin (day 42). Germline variant calling was limited to single nucleotide variants (SNVs) and small insertions and deletions (indels).

Germline SNVs consisted of the union of variant calls from two variant callers, Samtools [5] and VarScan [6]. Samtools variants were called using 'samtools mpileup' with parameters '-BuDS' followed by filtering with the GMS filter tools 'var-filter-snv' and 'false-positive-vcf v1' with parameters '--max-mm-qualsum-diff 100 --bam-readcount-version 0.4 --bam-readcount-min-base-quality 15'. VarScan variants were generated using version 2.3.6 of the software with parameters '--nobaq --min-coverage 3 --min-var-freq 0.20 --p-value 0.10 --strand-filter 1 --map-quality 10'. The resulting SNVs were filtered using the GMS filter tool 'false-positive v1' with parameters '--max-mm-qualsum-diff 100 --bam-readcount-version 0.4 --bam-readcount-min-base-quality 15'.

Germline indels were called using VarScan (version 2.3.6) with parameters '--nobaq --min-coverage 3 --min-var-freq 0.20 --p-value 0.10 --strand-filter 1 --map-quality 10'. The resulting indels were filtered using the GMS filter tool 'false-indel v1' with parameters '--max-mm-qualsum-diff 100 --bam-readcount-version 0.4 --bam-readcount-min-base-quality 15'. Refer to Griffith et al. 2015 for detailed descriptions of the SNV and indel filters [2].

The predicted functional relevance of SNVs and indels were assessed by annotation with the Ensembl Variant Effect Predictor (VEP) [7] and Gemini (v0.16.3) [8]. The Gemini query was constructed to limit results to coding variants only and remove all variants that are reported as common polymorphisms in the 1000 genomes or the exome sequencing project. The list of candidates was further limited to only those variants occurring within genes of the Cancer Gene Census [9]. All variants in this final list were manually reviewed. Variants within regions with suspected but unassembled pseudogenes or other partial gene segments missing from the human reference sequence, such as *PDE4DIP*, *USP6*, *NOTCH2*, *MAML2*, and *FANCD2* were removed from the analysis [10]. Passing variants were re-annotated with the GMS annotator [2] and transcript models obtained from Ensembl version 74 (GRCh37) (**Table S3**).

Somatic variant analysis

Somatic variant calling was performed using tools optimized for each variant type that were integrated into the GMS [2] (refer to **Tables S3**, **S4**, **S6** and **S7** for all final variant calls). For this case we used GMS processing profiles 'January 2015 WGS Somatic Variation' and 'January 2015 Exome Somatic Variation' for exome and WGS data, respectively. For all variant types, multiple variant calling and filtering algorithms were employed and combined to produce a final candidate set. The majority of these calls were subjected to validation by comparison across the primary datasets (WGS, exome and/or RNA-seq). Most SNVs and CNVs were also subjected to validation by DNA capture and deep sequencing (described below). All variants provided in the supplementary materials of this manuscript were subjected to manual review except for non-coding SNV lists in **Tables S3**, **S6**, and **S7**. Visualizations of support for all individual variants mentioned by name in the manuscript are provided in the supplementary materials.

Single nucleotide variant candidates were identified by a combination of five somatic SNV callers: Samtools [5], SomaticSniper [11], VarScan [6], Strelka [12], and Mutect [13] as previously described [1]. To produce each individual call set, each variant caller utilized custom parameters and filtering. The filtered Samtools calls were intersected with SomaticSniper calls and variants found by both these callers were joined by union with calls from VarScan, Strelka, and Mutect. Samtools variant calls were produced with 'samtools mpileup' using the parameters '-BuDS'. These Samtools calls were filtered by the GMS filter 'false-positive-vcf v1' with the following parameters '--max-mm-qualsum-diff 100 --bam-readcount-version 0.4 --bam-readcount-min-base-quality 15'. Variants were called by SomaticSniper (version 1.0.4) using the parameters '-F vcf -G -L -q 1 -Q 15'. These SomaticSniper calls were filtered using the GMS filter tool 'false-positive v1' with parameters '--bam-readcount-version 0.4 --bam-readcount-min-base-quality 15' and the GMS filter tool 'somatic-score-mapping-quality v1' with parameters '--min-mapping-quality 40 --min-somatic-score 40]'. VarScan variant calls were produced with version 2.3.6 of the software using default parameters except '--nobaq'. The resulting variants were filtered using GMS filter tools 'varscan-high-confidence v1' and 'false-positive v1' with parameters '--bam-readcount-version 0.4 --bam-readcount-min-base-quality 15'. Strelka variant calls were generated using version 1.0.11 with default parameters except 'isSkipDepthFilters = 0' was used for WGS data and 'isSkipDepthFilters = 0' was used for exome data. Mutect variant calls were produced with version 1.1.4 with default parameters. Mutect analysis was performed in parallel on 50 approximately equally sized subsets of the reference genome sequence. Mutect calling was guided by known mutation positions from Cosmic version 54 [14] and known SNPs from dbSNP version 137 [15].

Small somatic insertions and deletions were identified by a combination of four somatic indel callers: GATK, Pindel [16], VarScan, and Strelka as previously described [1]. To produce each individual call set, each variant caller utilized custom parameters and filtering. The indel final call set was the union of GATK, Pindel, VarScan and Strelka. GATK indel calling used the now deprecated GATK indel caller with default parameters. Pindel v0.2.2 was run with the parameter '-w 10' with a config file generated to pass both tumor and normal BAM files set to an insert size of 400. Pindel results were filtered using the GMS filters 'pindel-somatic-calls v1', 'pindel-vaf-filter v1' with parameters '--variant-freq-cutoff=0.08' and 'pindel-read-support v1'. VarScan indel calling was performed using version 2.3.6 of the software with default parameters except for '--nobaq'. VarScan variants were filtered by the GMS filter tool 'varscan-high-confidence-indel v1'. Finally indels were called using Strelka version 1.0.11 with default parameters except 'isSkipDepthFilters = 0' and 'isSkipDepthFilters = 1' were used for the WGS and exome data respectively.

All transcript variant annotations (e.g. predicted amino acid effects) for SNVs and indels were generated by the GMS's variant annotator [2] using transcript models obtained from Ensembl version 74 [17]. Variants were also annotated where appropriate with information from dbSNP (version 138) [15].

Copy number variants (CNVs; large-scale amplifications or deletions) were identified by observing differences in WGS coverage between tumor samples and the normal skin reference sample. Coverage was calculated for fixed size windows of the reference genome using the GMS tool 'bam-window' version 0.5 with parameters '-w 10000 -r -l -s -q 1'. The resulting window coverage values were analyzed by CopyCat (https://github.com/chrisamiller/copyCat) with parameters '--per-read-length --per-library'. CopyCat corrects window values to account for GC bias, mappability, and other factors. The resulting window values were segmented by a circular binary segmentation algorithm to identify series of consecutive windows that may represent individual CNV events. CNVs were predicted from Exome data using the R package cn.mops [18]. CNVs were predicted from genotype microarray data by use of circular binary segmentation on normalized microarray intensity values corresponding to each SNP position assayed. All copy number variants were visualized and manually reviewed using the GMS tool 'CnView' and custom R scripts.

SVs were identified by a custom pipeline in the GMS essentially as previously described [1]. Briefly, the tool BreakDancer (version 1.4.5) was used with parameters '-g -h:-a -t -q 4 -d'. The resulting candidates were filtered via a GMS tool utilizing NovoCraft NovoAlign alignment of read assemblies created by Tigra [19]. SVs were also called with Manta (https://github.com/Illumina/manta) using default parameters. It produced a VCF (v4.1) file containing 367 candidate somatic structural breakpoints corresponding to translocations, inversions, deletions, insertions, and tandem duplications. 132 events passed the default filtering steps of Manta. All translocations were manually reviewed in IGV to determine potential genes affected. All translocations, deletions, and inversions were manually reviewed using both svviz [20] to realign reads to the predicted breakpoint sequence and IGV [21] to visualize coverage support. SVs passing this manual review are provided in **Table S3**.

The overall somatic mutation rate calculation for the second relapse tumor was determined by identifying the number of single nucleotide variants present in non repetitive regions of the genome that were successfully validated by custom capture and deep sequencing. To obtain a rate per Mb, this number was divided by the effective discovery space for these variants and multiplied by 1 million. The effective discovery space was the number of positions in the genome with sufficient coverage (>= 20x depth, base quality >= 20 and mapping quality >= 20) after excluding repetitive regions. The size of the effective discovery space was determined to be 1,298,325,076 bp.

One SNV/SNP position (C/T at 2:221,177,686) was filtered in a custom way. A somatic SNV was called at this position in the patient's second relapse WGS data (the sorted blasts sample). Careful examination of all data revealed that the patient's germline genotype at this position was C/C, the patient's tumor harbored a somatic C/T, the sibling allograft genotype was C/C, and by coincidence the second (MUD) allograft genotype was C/T. In other words, the patient by chance acquired a somatic SNV (C/T) at the site of a known polymorphism but where the patient was homozygous reference for the polymorphism (C/C). The patient's sibling was also homozygous (C/C) but the second allograft donor was heterozygous (C/T). This created a confusing pattern of transmission for this single variant over the time points sequenced. This variant was therefore removed from **Figure 4** and **Figure S10**, but only for the time points sequenced after the MUD allograft (day 3,219 and day 4,024).

Design and analysis of a custom capture array and validation sequencing
A custom reagent of biotinylated DNA oligonucleotides was designed and ordered from Roche NimbleGen to allow capture and deep sequencing of candidate regions. These candidate regions consisted of four categories: (A) somatic SNVs identified in the second relapse by analysis of the WGS and exome data, (B) SNPs found to be heterozygous by analysis of the WGS data from the skin normal that also lie within large-scale deletions found by analysis of the WGS data for the second relapse, (C) heterozygous SNPs from the normal skin data that occur in control (copy neutral) regions according to the second relapse data, and (D) all SNPs and exons of the *FLT3* locus. A total of 5,628 regions were submitted for probe design. These regions included 2,403 candidate somatic SNVs, 2,429 SNP heterozygous sites tiled across large-scale deletions, 536 heterozygous SNP sites tiled across control regions that were not copy altered, and 260 SNPs and exons of the *FLT3* locus. A total of 27 large-scale deletion regions and 285 copy neutral control regions were tiled. SNPs in the deletion and control regions were required to have a VAF in the normal skin sample between 40% and 60%. Deletion regions were required to have a copy number difference between tumor and normal that was at least -0.25. Control regions were required to have a copy number difference greater than -0.1 and less than 0.1. SNPs determined to be heterozygous in the normal skin WGS data were required to have a minimum coverage of 20x to be included. The custom capture reagent was applied to eleven samples from eight time points during disease progression and sequencing achieved ~800-900x coverage for the targeted regions (**Table S2**, **Figure 1**). The purpose of this array was to validate somatic SNVs and deletions and to obtain accurate VAFs for validated somatic SNVs. These VAFs would be used to monitor disease burden and examine clonal architecture during disease progression (**Figure 4**). SNVs initially detected by WGS or exome were considered validated as somatic if they were adequately supported by the deep custom capture data. Specifically, we required that this custom capture data provide: (A) >50x coverage in both tumor (SB_d3072_A) and normal skin samples, (B) >0% VAF in the tumor sample, and (C) <1% VAF in the normal skin sample. This resulted in 1,921 high quality validated somatic variants. The mutations spectrum analysis and visualization depicted in **Figure S5** was created using the GenVisR package (https://github.com/griffithlab/GenVisR) with variants from the master SNVs list (**Table S7**). Filters were applied to these variants requiring 50x coverage for tumor (SB_d3072_A) and normal samples. A VAF filter requiring ≥ 20% (tumor) and ≤ 1% (normal) was applied leaving a total of 1,756 variants after removing indels. Expectations were calculated by randomly mutating the 1,756 variants passing filters with equal probability (⅓) 100 times and taking the average. The variance of this method was found to be 6.02e-5.

RNA-seq analysis - alignment, expression estimation and fusion detection
Analysis of RNA-seq data including read alignment, estimation of transcript abundance, and fusion detection were conducted essentially as previously described [22]. Briefly, RNA-seq analysis was conducted using the GMS rna-seq processing profile 'December 2014 OvationV2 RNA-seq'. RNA-seq reads were pre-processed with flexbar [23] version 229 with parameters '--adapter CTTTGTGTTTGA --adapter-trim-end LEFT --nono-length-dist --threads 4 --adapter-min-overlap 7 --max-uncalled 150 --min-readlength 25'. The resulting trimmed reads were aligned to the human reference genome (version GRCh37 also known as hg19) using the aligner TopHat version 2.0.8 [24]. TopHat alignments were performed with default parameters except '--bowtie-version=2.1.0'. Quality of the resulting RNA-seq BAM file was assessed by FASTQC version 0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and samstat version 1.08 [25]. Transcript abundances measure as Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values were generated by Cufflinks version 2.1.1 with parameters '--num-threads 4 --max-bundle-length 10000000 --max-bundle-frags 10000000'. Additionally, during transcript abundance estimation a mitochondrial and ribosomal RNAs were 'masked' (excluded) from consideration by Cufflinks using the '--mask-file' parameter and a GTF

file containing these transcripts. Mitochondrial and ribosomal transcripts were identified using the 'biotype' field in a transcripts GTF file obtained from Ensembl. RNA fusions were predicted by use of ChimeraScan version 0.4.6 [26] with parameters '-p 2 --bowtie-version=0.12.7 --reuse-bam=0 --total-frag-limit=5 --span-frag-limit=1 --fusion-partner-limit=3' and Integrate version 0.2.0 [27] with default parameters.

Outlier expression analysis.

Gene transcripts were identified as outliers through a multiple filter pipeline (**Figure 3A**). Transcripts without at least one control with an FPKM >= 0.5 were discarded. The control FPKM values were used to approximate normal distributions for each transcript, and z-scores were calculated for each transcript using the ALL1 FPKM values. These z-scores were used as a secondary filter by keeping only the top 1% of genes by z-score (**Table S3**). This list of putative driver genes were submitted to DGIdb [28] (http://dgidb.genome.wustl.edu/), filtering for candidates that were considered clinically actionable, unambiguous, and with known drug-gene interactions. To allow comparison of RNA expression values from ALL1 to healthy tissue (**Figure 3C**) we also obtained RNA-seq data from eight healthy donors [29] (dbGAP Accession: phs000159). In this previous study, bone marrow aspirates from healthy donors were separated into populations of CD14+ (monocytes), CD19+ (B-cells; CD33$^-$/CD19$^+$), CD3+ (T-cells; CD33$^-$/CD3$^+$), CD34+ (hematopoietic progenitors), PMN (neutrophils; CD33$^-$/CD15$^+$/CD16$^+$), and Pro (promyelocytes; CD14$^-$/CD15$^+$/CD16$^{low/-}$) cells using fluorescence-activated cell sorting as previously described [29]. Finally, comparisons of ALL1 microarray expression data from sorted bone marrow to microarray expression data from 207 B-cell leukemia bone marrow (N=131) or peripheral blood (N=76) samples with >80% blasts (**Figure 3D**) from Kang et al. 2010 (GEO Accession: GSE11877, [30]). The ALL1 sample was processed (in triplicate) using the same microarray platform (Affymetrix Human Genome U133 Plus 2.0 Arrays).

Definition of "founding" variants

To identify likely "founding" variants of the second relapse (SB_d3072_A) (those likely to be present in every tumor cell rather than a subclone) a high quality set of variants were selected as follows. First, readcounts and VAFs were obtained from the combined (exome, wgs, capture) sequence data using bam-readcount (0.7) for all somatic variants within the capture regions. To ensure the accuracy of VAF estimates for this analysis, filters were applied before clustering to limit the somatic variants to copy neutral and LOH-free regions (i.e. variants in deletion regions were excluded). Similarly, variants with a variant allele frequency greater than 5% in the normal skin (day 42) sample were removed. A coverage cutoff was also applied requiring a variant to be covered at or above the second decile in both the normal and tumor sample effectively removing variants within an allosome and ensuring adequate depth. 1,588 variants remained after filtering with an average coverage of 1,293x in the sorted blast sample ('SB_d3072_A'). A k-means clustering algorithm was used to identify clusters of somatic variants with similar VAF (using k=3, n=100). Variants within the highest density cluster centered around 50% VAF were classified as founding with remaining clusters designated as subclones (**Figure S9-S10**). To determine the frequency of founding variants originally identified in the second relapse sample ('SB_d3072_A') at different time points, variants with such a designation were extracted for 12 samples. A variant was considered for this analysis if the variant had ≥ 20x coverage for each individual sample. Additionally, VAF filters unique to each sample, mirroring those filters used for the purity estimate, were applied (**Table S1**, see below for more details on purity estimation). The number of founding variants from the second relapse samples detected in each earlier or later sample are provided in **Table S1**.

Tumor purity, contamination, and clonal architecture

An estimate of both tumor purity and sample contamination, defined as the proportion of tumor cells within a sample and the proportion of cells contributed by allografts received by the patient was performed. These estimates were derived from analysis of VAFs obtained after combining all sequence data available (WGS, exome, and capture) for each sample to create a single master variant read counts table (**Table S7**).

To estimate tumor purity, founding variants were first identified in the sorted blasts sample from relapse 2 (described above). To ensure sufficient depth, variants were required to have >= 20X coverage within each sample. VAF cutoffs specific to each sample (ranging from 1-10%) were applied to ensure that variants acquired after the time point analyzed were not included in purity calculations (**Figure S10**). Purity estimates were calculated for each sample by multiplying the VAFs by two and taking the median for each sample.
To estimate potential contamination of each patient DNA sample by DNA from the sibling allograft donor, we analyzed 536 heterozygous control SNPs on the custom capture panel (see detailed custom capture description above) and used three samples as reference points. Heterozygous SNPs in the patient were determined by examination of the skin normal sample obtained at day 42 during the first remission. Since these variants were heterozygous in the ALL1 patient we expected a VAF close to 50%. In samples obtained after an allograft, we expect deviation from 50% for a subset of SNPs that are homozygous in the donor. To identify SNPs homozygous in the DNA of the allograft donor we examined the sorted lymphs sample, where cells corresponding to the sibling allograft donor were purified as a control (see detailed sample description above). The degree of deviation from the expected 50% for these variants was used as a measure of the percent contamination of each patient sample by the donor's bone marrow cells (**Figure S6**). To calculate these estimates, VAFs for the control SNPs were extracted from each BAM file and a coverage filter was applied requiring a position to be covered above the second decile for each time point. Variants in CNV/LOH regions or within an allosome were removed for this analysis leaving 414 heterozygous control SNPs. Heterozygous variants (60 ≥ x ≥ 40 % VAF) in the normal that were homozygous after the first allograft according to the sorted lymph sample (x < 40 or x > 60 % VAF) were selected and contamination rates calculated (contamination rate = |variant position - 50| x 2). Variants that were heterozygous in the normal sample, heterozygous in the sorted lymph sample, but homozygous after the second allograft were selected and contamination rates calculated for the second allograft. The median of these calculations is reported in **Table S1**. Note that since all samples with the exception of "M_d3219_I" and "BM_d4024_I" were obtained prior to the second allograft, we expect no contamination from this genotype prior to day 3,219. This second estimate was performed to gauge the measurement error of the approach.

The clonality analysis depicted in **Figure S9** was performed using a modification of the approach described in Miller et al. 2014 [31]. Briefly, variants targeted by the custom capture sequencing described above were obtained and limited to those that were somatic and supported by both WGS and custom capture data. Filters were applied requiring that: (A) the VAF in the normal skin sample be less than 5%, (B) normal coverage was greater than the 2nd decile (867x) and (C) tumor coverage was greater than the 2nd decile (958.4x). Variants within LOH or CNV regions were removed and a clustering algorithm was applied (K-means, k=3, trials=100). Variants plotted were ceilinged at a coverage of 2,000x (i.e. set equal to 2,000x for display purposes in this plot).

Neo-epitope prediction

We used a pipeline 'pVAC-Seq' (https://github.com/griffithlab/pVAC-Seq) developed at the McDonnell Genome Institute to assess candidate epitopes resulting from somatic mutations. Briefly, each of the coding missense mutations identified in ALL1 was annotated with predicted amino acid changes, which were then translated into a 21-mer amino acid FASTA sequence, with ideally 10 amino acids flanking on each side of the mutated amino acid. This FASTA was then evaluated through the HLA class I peptide binding algorithm NetMHC 3.4 [32, 33] to predict high affinity dextramer (10mer) peptides for the mutant as well as the wild type. This was done for all five patient alleles- HLA-A02:01, HLA-A31:01, HLA-B08:01, HLA-B18:01 and HLA-C07:01 and differences in binding affinities were calculated. The candidate neoepitope peptides were filtered to include only those with binding affinity IC50 value < 500nm and the best representative neoepitopes per mutation across all alleles were considered for further analysis (**Table S3**).

Cloning and sequencing of full length *EP300-ZNF384* clones

To verify the structure of the *EP300-ZNF384* fusion (**Table 1**) predicted in the whole genome data by Breakdancer [34] and Manta (https://github.com/Illumina/manta) and transcriptome data by ChimeraScan [26] and Integrate [27], we performed full length cDNA cloning and sequencing to capture the fusion sequence. Amplicons for cloning were generated from cDNA obtained from the second relapse (sample 'SB_d3072_A_RNA') using primers targeting the 5' UTR of EP300 and 3' UTR of *ZNF384* as follows: *EP300* 5'UTR (5'-GCGAATTTGTGCTCTTGTGC-3') and *ZNF384* 3'UTR (5'-CCTGTGAAGGAAAGCCGTGA-3'). Amplicons were ligated into the cloning vector PCR2.1 and were subjected to a primer walking sequencing strategy on an ABI Sanger sequencing instrument. A total of 25 sequencing primers were used. Sequence assemblies for each clone were created by use of a Phred/Phrap/Consed pipeline [35-37] followed by manual sequencing finishing. The resulting consensus sequence was exported for each clone to determine the precise cDNA breakpoint of the *EP300-ZNF384* and to confirm the presence of two somatic missense mutations that were found to be in linkage with the fusion event. Two *EP300-ZNF384* clones were sequenced in this manner. These clones agreed exactly on the sequence of the breakpoint and the presence of two somatic missense mutations in the *EP300* portion of the fusion. These two somatic mutations were also observed in the WGS, Exome and RNA-seq data from the second relapse tumor. Each clone also contained three additional differences that were not shared between the two clones, thus representing either sequencing or cloning artifacts. To obtain a clone without artifacts we produced indexed libraries for an additional 34 clones and sequenced them on an Illumina MiSeq instrument. The resulting reads were aligned against a custom reference sequence consisting of the expected *EP300-ZNF384* sequence based on the human reference genome and our knowledge of the cDNA breakpoint. All 34 clones shared the same fusion breakpoint and the two somatic missense mutations observed in the first two clones. These clones had between 1 and 9 artifacts each. One clone 'MC15b' had only a single base change in the last codon of the open reading frame of the fusion and this mutation was conservative in that it maintained the stop codon (TAG -> TAA, Stop -> Stop). This clone was used for functional experiments.

Personalized assays for disease monitoring

Three personalized genomic assays were developed to monitor disease in the patient. These were based on three types of somatic events: single nucleotide variants (SNVs), large-scale deletions, and the *EP300-ZNF384* structural variant. To monitor disease burden by use of somatic SNVs we started with 1,921 high-quality validated variants discussed in the custom capture and validation sequencing section above. These variants were further filtered down to 1,588 variants by limiting variants to those passing filters identical to those used in **Figure S9** (**Methods**). The VAFs for remaining variants were calculated by combining all sequence data and plotted for each disease time point using the SinaPlot method (https://github.com/sidiropoulos/sinaplot). The primary (day 0 clot and slide) samples had an additional 40x coverage filter applied to compensate for the low depth compared to the capture time-point data leaving 270 and 581 variants, respectively (**Figure 4A**). The same approach was used to compare the final time point subjected to deep capture sequencing (day 3,219) to the persistent relapse sample (day 3,107) and disease-free normal skin sample (day 42) (**Figure 4C**).

To assay tumor burden by quantitative FISH, four deletions were assayed by interphase nuclei FISH using custom probes for a 5 Mb deletion on chromosome 12 (12p13.2 - 12p12.3), a 15 Mb deletion on chromosome 18 (18p11.32 - 18p11.1), a 26 Mb deletion on chromosome 20 (20p13 - 20p11.1) and a 39 Mb deletion on chromosome 9 (9p24.3 - 9p13.1). Refer to **Table S3** for additional details on each of these deletion regions. 200-1000 nuclei were assessed for each time point in the quantitative interphase nuclei FISH assay and the percent of cells were recorded and plotted for each time point examined (**Figure 4B**).

Finally, to assay the presence of *EP300-ZNF384* throughout progression of the tumor (**Figure 4D**) we performed qualitative PCR using primers that flanked the genomic fusion breakpoint and were expected to produce a 168 bp product from genomic DNA. This assay was applied to genomic DNA isolated from the day

42 normal skin ('Skin_d42_I'), day 0 marrow core ('MC_d0_clot_A'), day 42 marrow ('BM_d42_I'), day 1,893 relapse 1 ('M_d1893.1_A'), and day 3,072 relapse 2 ('SB_d3072_A') samples. Primer sequences for the EP300-ZNF384 DNA breakpoint were as follows, Left Primer 5'-CTAGAGTAACAGGGACCAAAGAGTA-3' (targets EP300 side of breakpoint), Right Primer 5'-GACCCACACATGCATCAAAACA-3' (targets ZNF384 side of breakpoint). To further assess the presence of EP300-ZNF384 throughout progression of the tumor we performed quantitative PCR (qPCR) using primers that flanked the genomic DNA fusion breakpoint. This assay was applied to genomic DNA obtained from the day 42 normal skin ('Skin_d42_I'), day 42 marrow ('BM_d42_I'), day 1,893 relapse 1 ('M_d1893.1_A'), and day 3,072 relapse 2 ('SB_d3072_A') samples, and a sample obtained after the final remission. The same primers used for PCR described above were used for this qPCR assay. Positive control primers for the qPCR targeted *HBB* (Beta Globin) with sequences as follows: Left Primer 5'-CTAATGCCCTGGCCCACAAG-3', Right Primer 5'-AGATGCTCAAGGCCCTTCATA-3'. A series of standards was created using genomic DNA from the second relapse (sorted blasts) sample diluted to 1:1, 1:10, 1:100, 1:1000, and 1:10000. The 1:1 sample was used as a positive control in each run. Negative controls consisted of water, genomic DNA from the skin obtained from the patient during remission (day 42), and skin from an unrelated sample. The values shown in **Figure 4D** were calculated as follows: ($\sqrt{2^{-\Delta\Delta CT}}$). First a delta cycle threshold (delta CT) value was calculated for each replicate by subtracting the Beta Globin CT value from each sample CT value. Next a delta-delta CT value was calculated by subtracting the delta CT value for the 1:1 positive control from the sample delta CT value. A more intuitive abundance value was then calculated as $2^{\text{-(delta-delta CT value)}}$. The geometric mean of three such values for each sample (technical replicates) was calculated. Error bars display the 95% confidence interval (assuming normal distribution) for each mean. Finally, all values were subjected to a square root normalization for display purposes.

Allele specific assay of *FLT3* transcript expression using a heterozygous *FLT3* SNP

To assay allele-specific mRNA expression, total RNA was purified from aspirate smears from archived bone marrow samples. The aspirate smear coverslips were removed by soaking overnight in xylene. Tissue was scraped off coverslips with clean razor blades into microcentrifuge tubes, and then washed and pelleted in 70% ethanol to remove the xylene. Pellets were resuspended in 100ul RNA lysis buffer from Quick-RNA MicroPrep Kit (Zymo Research, Irvine, CA), and purified according to kit instructions. cDNA libraries were made from each sample using Quantitect Reverse Transcription Kit (Qiagen, Valencia, CA) and subjected to 40 cycles of PCR (qPCRBIO SyGreen Hi-Rox, PCRBiosystems, St. Louis, MO) using primers designed to amplify a 188bp amplicon that included a heterozygous C/T single nucleotide polymorphism (Left primer 5'-AATGGGTGCTTTGCGATTCA-3', Right Primer 5'- CAATGTGGTCTGAGGAGTTTGA-3'). This SNP is located at cDNA position +602 from the transcriptional start site (dbSNP version 142 ID: rs1933437). In genomic coordinates the SNP is at '13:28,624,294(G/A)' (reference build GRCh37). To obtain digital read counts for each allele, a second round of PCR was performed using barcoded versions of the same primers used in the first round. PCR product from each sample was purified using Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN), pooled and sequenced on the Ion PGM System (Thermo Scientific, Waltham, MA).

Cloning and sequencing of *FLT3* intron 1 indel mutations

To validate the genomic sequence of the *FLT3* intron 1 indel mutations (**Figure S40**) PCR amplicons flanking each indel were generated, cloned and sequenced. The primers used to generate these amplicons were as follows: Indel1 Left Primer 5'-TGGAAATTCCCAGAATCCAG-3', Indel1 Right Primer 5'-GGGCCCAAAGAGGATAAATG-3', Indel2 Left Primer 5'-CATGGTGGACAGCACCTG-3', and Indel2 Right Primer 5'-GACACTGGAGGTTTGCCACT-3'. Amplicons were ligated into the cloning vector PCR2.1. A total of 8 clones were generated, one for each indel using template DNA from the relapse 2 tumor, and 3 for each indel using template DNA from the normal skin. Each clone was subjected to a primer walking sequencing strategy on an ABI Sanger sequencing instrument. A total of 8 sequencing primers were used. Sequence assemblies for each clone were created by use of a Phred/Phrap/Consed pipeline [35-37] followed by manual sequencing finishing. In addition to Sanger sequencing, additional amplicons for Indel1 and Indel2 were generated from multiple disease time points, indexed, pooled and sequenced on an Illumina MiSeq instrument. The two outermost primers (Indel1 Left and Indel2 Right) were also used to generate an amplicon of ~1,200 bp that would contain both *FLT3* indel 1 mutations if they were in linkage (they were not).

Integration of WGS, Exome and RNA-seq data and clinical interpretation
To produce a final report for the ALL1 case and generate starting points for the figures of this manuscript we used the GMS "clin-seq" (aka "med-seq") pipeline to integrate results from the McDonnell Genome Institute's automated pipelines for read alignment, somatic variant detection, and RNA-seq analysis [2]. This pipeline attempts to cross-validate variant calls from multiple variant calling algorithms and complementary data sets such as WGS and exome. For example, fusion candidates identified in the RNA data were intersected with interchromosomal translocation candidates identified in the genomic DNA data. A detailed final report of SNVs and indels was generated to place all variants in the context of known polymorphism data from dbSNP and 1000 genomes and to aggregate read support across all sequence datasets generated for this work (**Table S3**). To assess the expression status of SNVs and small indels identified in the genomic DNA, readcounts and VAFs were generated for these site from the RNA-seq BAM file using bam-readcount (https://github.com/genome/bam-readcount). All single nucleotide variants and small indels were compared to Cosmic [14] to identify mutations occurring in known hotspots. For further details on the integrative analysis performed by the clin-seq pipeline, refer to Griffith et al. 2015 [2]. To assist in the interpretation of biological and clinical relevance all affected genes were annotated with results from the Drug Gene Interaction database (DGIdb) [28] (http://dgidb.genome.wustl.edu/), and all variants were used to query the Database of Canonical Mutations (DoCM) (http://docm.genome.wustl.edu/) and Clinical Interpretation of Variants in Cancer (CIViC) resource (https://civic.genome.wustl.edu/).

SUPPLEMENTARY RESULTS

The first individual blood cancer [38] and solid tumor [39] whole genomes were sequenced and published within the last several years and were followed by extensive surveys of cancer exomes (and some whole genomes) by The Cancer Genome Atlas (TCGA) [40] and International Cancer Genome Consortium (ICGC) [41]. These efforts (and also sequencing focused on integrating exomes and transcriptomes) have considerably expanded our understanding of the recurrent mutations, structural variants, and copy number alterations that occur in leukemia cohorts [42-45].

Genome analysis
The patient (hereafter referred to as ALL1) was appropriately consented for whole genome sequencing on an IRB approved protocol. We performed a combination of genotype microarray analysis, expression microarray analysis, whole genome sequencing (WGS), whole exome sequencing, transcriptome sequencing, custom capture sequencing, qPCR, and quantitative interphase FISH. Although the immediate analysis focused on the second relapse sample, we were able to ultimately obtain 18 samples from 9 time points throughout disease progression to better understand the evolution of this tumor over time (**Figure 1A**, **Table S1**). The specific advantages of each technology were leveraged to create a comprehensive combinatorial analysis of the tumor, leading to the development of custom disease monitoring methods, which continue to be employed in disease management of the patient (**Figure 1B**). Exome and whole genome sequencing were performed for a matched skin normal, two leukemic bone marrow samples from the original presentation, the first relapse marrow, and the second relapse marrow. Both original tumor samples were fixed archival samples: one was from a bone marrow clot section and the other obtained from a fixed and decalcified bone marrow biopsy section. Both yielded heavily degraded DNA. The first relapse sample contained a low percentage of tumor cells. The second relapse (post-allogeneic transplantation) was a fresh sample from which blasts were sorted so that

contaminating normal cells from the transplant donor would not confound the analysis (CD45 dim, low side scatter, CD19+/CD34+) (**Figure S2**, **Methods**). Exome sequencing of these samples achieved 179-264x coverage and whole genome sequencing achieved 30-66x coverage (**Table S2**, **Figure S3**). WGS and exome data from all tumor samples were analyzed to exclude the possibility of contamination or sample swaps (**Figure S4**). The majority of *de novo* variant discovery was performed using genomic DNA obtained from the sorted second relapse sample. The status of these individual variants was then assessed retrospectively at the earlier time points. Based on the detection and validation of 1,921 somatic variants, we estimated a mutation frequency of ~1.48/Mb in the second relapse sample. The genome of the second relapse exhibited enrichment for single nucleotide variants (SNVs) that are transitions (**Figure S5**), and a substantial number of small insertions and deletions (indels), large-scale deletions, translocations, and other structural variants (**Table S3**).

Analysis began with the search for clinically relevant somatic mutations in the whole genome and exome data of the second relapse. Variant candidates identified in these data were used to design a custom capture reagent for 5,628 regions of interest (**Methods**). This capture reagent was used to perform deep (~800-900x) validation sequencing for variants discovered by WGS or exome, and later, to assess the presence of each variant at additional samples/time points not previously sequenced (**Table S2**). Additional regions of interest were used to assess both large-scale deletions and sample quality and purity. Since the second relapse sample was obtained after the patient received a matched sibling allograft, we were concerned that some somatic variants identified in the second relapse might actually represent contamination from the sibling donor cells despite our use of sorted blasts for the analysis (**Figure S3**). To assess this issue we included hundreds of SNPs in our custom capture reagent that were found to be heterozygous in the normal skin sample obtained from the patient. Analysis of these variants demonstrated that somatic variants observed in the sorted blasts from the second relapse at a VAF greater than ~0.5% were unlikely to be artifacts of contamination (**Methods**, **Figure S6, Table S1**). Of the 2,403 candidate somatic SNVs included in the custom capture design, 2,339 (97.3%) achieved ≥ 50x coverage in both the normal skin and sorted blasts samples and of those 1,931 (80.4%) were validated as somatic (**Methods**). We also defined a set of high quality variants (**Methods**) that excluded SNVs within regions of copy number variation (CNV) or loss of heterozygosity (LOH) (**Figure S7-S8**). The resulting 1,588 validated SNVs in copy neutral regions were used to model the clonal architecture of the second relapse tumor: a founding clone and at least two subclones were apparent (**Figure S9**). The dominant cluster of high quality VAFs identified in the second relapse was used to estimate tumor purity, resulting in purity estimates of ~85% for the primary sample, ~16% for the first relapse sample, and ~94% for sorted blasts from the second relapse (**Methods**, **Table S1**, **Figure S10**). Extending our clonality analysis to the additional unsorted sample obtained at second relapse (day 3,072), we observed at least three subclones, suggesting that the sorting of blasts resulted in clonal skewing that obscured one subclone (**Figure S10**). Based on pairwise comparisons of VAFs observed at primary, first relapse, and second relapse, we identified events that persisted throughout the course of the disease, suggesting that they were present in the original founding clone. We also observed variants that were lost during disease progression, and variants that were acquired or enriched during progression (**Figure S11-S12**). Based on these observations, we infer that at least six distinct subclones existed at some point during the disease. While insufficient to create a complete model of clonal evolution [1], we have sufficient data to conclude that a dramatic shift in clonal architecture occurred between the initial presentation and the second relapse. For example, of 898 validated mutations in the dominant clone of the second relapse (i.e. present in every tumor cell of that sample), only 273 (30.4%) were detectable in the primary “clot” tumor despite adequate coverage (≥ 20x coverage) for 723 of these mutations (**Table S1**, **Figure S10**, **Methods**). Most variants that appear to be ‘gained’ in the relapse were likely present in the primary but were present within a low frequency subclone that below our limit of detection [1]. This gain and loss of subclones is consistent with a recent report evaluating the clonal architecture of pediatric ALL [46].

Potentially relevant somatic variants were identified by integrated analysis of all data using multiple algorithms for detection, annotation and visualization of each type of variant (**Table 1, Table S3** and **Methods**).

Germline variant analysis was also performed, but did not reveal any established candidates for susceptibility to B-ALL (**Methods**, **Table S4**). A total of 98 somatic SNVs with predicted coding effects were identified across the primary and two relapses. Among these were two missense mutations in *EP300* (P250H and P252S) (**Figure S13**) and a nonsense mutation within *NF1* (R2258*) (**Figure S14**). The *EP300* mutations are in linkage with each other and were present at a high variant allele frequency (VAF) in the primary sample and in both relapses. The *NF1* nonsense mutation, previously observed three times in the COSMIC database [14], appeared to have been acquired or enriched between the initial presentation and the first relapse, and exhibited the highest VAF of all coding mutations in the second relapse (**Table S3**). Among nine validated small insertions or deletions (indels), a 14-bp frameshift mutation was identified in *SETD2* (**Figure S15**), a tumor suppressor that influences chromatin state and transcriptional elongation.

In the second relapse sample, we observed 25 regions of copy number alteration including large-scale deletions as well as focal amplifications and deletions from the whole genome sequence analysis (**Methods**, **Table 1**, **Table S3** and **Figures S16-S31**). The majority of deletion events were supported by corresponding evidence for loss of heterozygosity (LOH). We observed only a single copy neutral LOH event affecting most of chromosome 17 (**Figure S27**). Based on the magnitude of LOH observed, this event appears subclonal and is the likely explanation for the higher than 50% VAFs observed for the *NF1* R2258* mutation. Focal events defined as < 5 Mb (0.2 - 4.8 Mb) included deletions affecting *SETD2* (3p21.31, **Figure S17**) and *IKZF1* (also known as *Ikaros*, 7p12.2 - 7p12.1, **Figure S20**). The *IKZF1* deletion is approximately 70 Kb in size, affects exons 4-8 (of reference transcript NM_006060) consequently removing the N-terminal end of the protein, and possibly giving rise to a dominant negative form previously described in childhood ALL [47]. This region of *IKZF1* is unusual: near the deletion breakpoint, we observed an assembly gap in the reference genome sequence build (GRCh37). It is possible that there is a complex repetitive element or segmental duplication that might be related to instability and focal deletion at this site. Several large-scale deletions (> 5 Mb) were also detected, including one affecting the tumor suppressor *RB1* (13q14.13 - 13q14.3, **Figure S26**) and another affecting *ETV6* (aka *TEL)* and *CDKN1B* (12p13.2-12p12.3, **Figure S24**). A large complex deletion associated with multiple breakpoints and rearrangements was observed on 10q (10q24.1-10q24.33, **Figure S22**) that affects multiple cancer genes including *TLX1*, *NFKB2*, *SUFU*, and *NT5C2*. Based on the coverage depth of each CNV event, we estimated whether it was likely to be heterozygous (single copy loss), homozygous (two copy loss), or heterozygous in a sub-clonal population of the tumor (**Table S3**) (with the caveat that it would be difficult to tell from the WGS CNV and LOH data the difference between a two copy deletion in a sub-clone, and a single copy deletion in the founding clone). Based on these estimates it seemed likely that the focal chromosome 7 deletion (*IKZF1*, **Figure S20**), and chromosome 12 (*CDKN1B* and *ETV6*, **Figure S24**) and chromosome 13 losses (*RB1*, **Figure S26**) were heterozygous and present in the dominant clone of the second relapse. In addition to the large-scale CNVs described above we identified smaller deletions and other structural variants (SVs) by analysis of read pairing and alignment data (**Methods**). For example, we detected a heterozygous ~13 Kb deletion of the first exon of the transcription factor *XBP1* (**Table 1**, **Figure S32**) and a 544 bp inversion of the 3rd exon of polymerase subunit *POLR2C*. Finally, SV analysis of the WGS data for second relapse predicted three translocations potentially resulting in fusion genes (**Table S3, Figure S33-S37**). One of these, resulting from a balanced translocation t(12;22)(p13;q13) (**Figure S33**) was predicted to result in an *EP300-ZNF384* fusion gene and protein (**Figure S34**). The genomic breakpoint of the fusion was validated by RT-PCR (**Figure S35A**), and qPCR of genomic DNA demonstrated that this translocation event was present in the primary tumor samples obtained at diagnosis (**Figure S35B**). For this reason, the EP300-ZNF384 fusion is a candidate initiating event for this tumor. This hypothesis is supported by a recent study that identified the same fusion in two B lymphoblastic leukemia cases [48]. Two additional gene fusions, *TBX19-SUFU* (**Figure S36**) and *ADCY10-CC2D2B* (**Figure S37**) were predicted from the WGS data but neither of these was predicted to result in an open reading frame. While the genome analysis described above led to the identification of several events relevant to the biology of this tumor (summarized in **Figure 2** and **Table 1**), none suggested treatment strategies once the patient was determined to have refractory disease following salvage therapy.

Transcriptome analysis

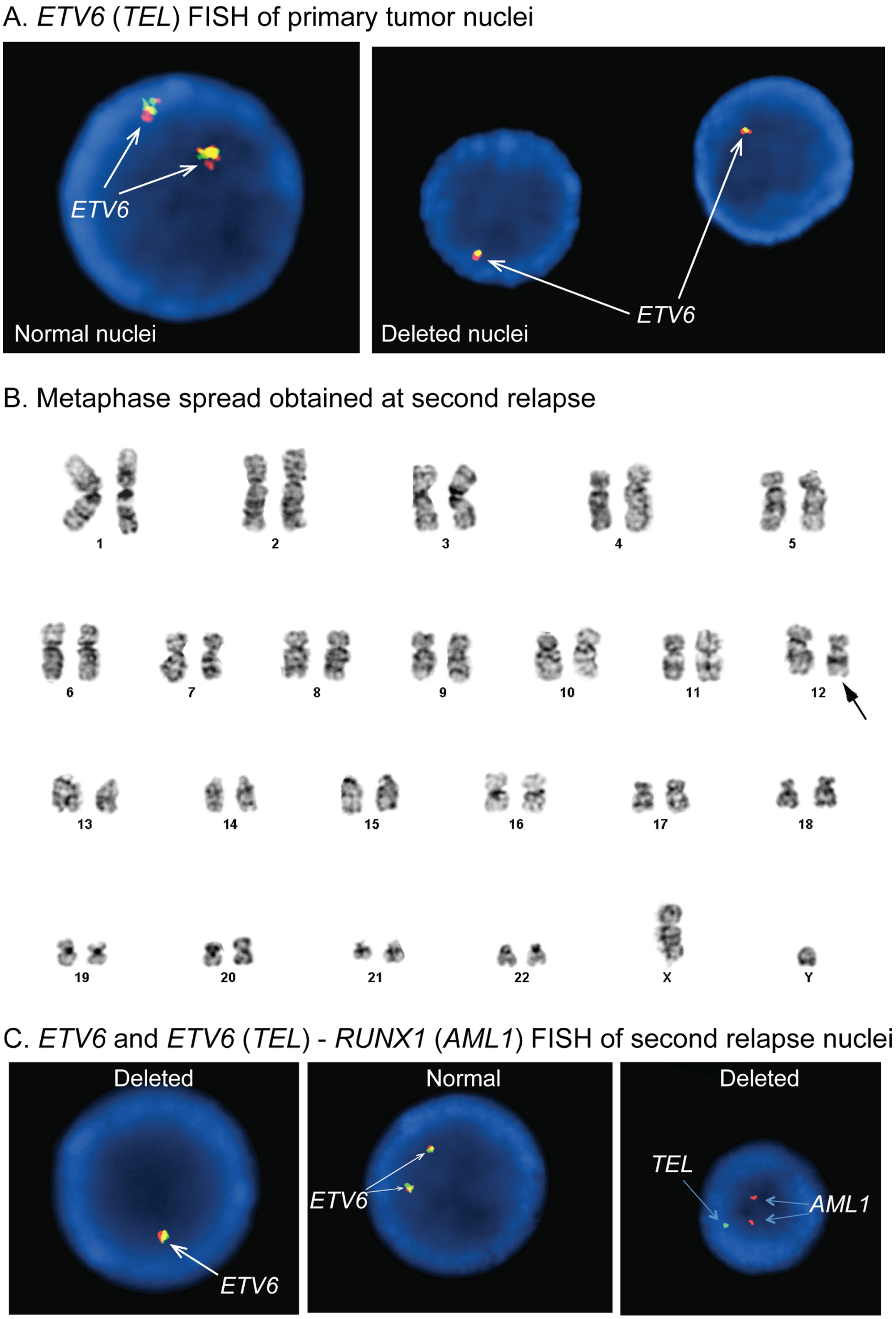
Several transcriptome analyses and methods for integration of RNA-seq data with WGS and exome data were developed during the course of this analysis. At the second relapse, 35 coding somatic SNVs were present and 62% of these were confirmed as expressed by RNA-seq (**Figure S38**). RNA-seq also confirmed expression of the *EP300-ZNF388* fusion (**Figure S39A**) described above. Full length cDNA cloning and Sanger sequencing established the complete fusion RNA sequence. This cDNA includes an ORF that maintains the 5' reading frame of *EP300* and 3' reading frame of *ZNF384*, including the entire C2H2 zinc finger domain (**Figure S34B**). Gene fusions with *ZNF384* as the 3' partner (but different 5' partners) have been previously reported in ALL and other leukemias [49-52] (**Figure S39B**). Fusions involving *EP300* have not been previously described. We used predictive analysis of microarray (PAM) method [53, 54] to establish that the patient was not of the 'Ph-like' subtype (**Methods**) associated with poor outcome, increased prevalence with age, and the presence of kinase-activating mutations that may be targetable [55]. Additional analysis using the 'ROSE' algorithm [56] clustered the patient with 'R5' cases that are enriched for fusions similar to the *EP300-ZNF384* event we observed.

Comparison of expression data from the second relapse to several sample cohorts was used to identify possible outlier expression of genes relevant to B-ALL treatment **(Methods, Figure 3**). Sample cohorts for comparison consisted of sorted blood cells from healthy donors, and 207 additional B-ALL tumors from pediatric patients. To identify outlier genes, a multiple-step filter approach was taken, selecting for differentially expressed, clinically actionable, and druggable genes using a combination of statistical thresholds and the Drug-Gene Interaction database [28] (**Methods, Figure 3A**). The resulting four potential gene targets were *FLT3, PDGFRB, WT1,* and *TUBB3* (**Table S3**). Of these, the fms-related tyrosine kinase, *FLT3*, had an estimated transcript abundance that was several orders of magnitude above the other three. Evaluation of the RNA expression estimates of *FLT3* in ALL1 and controls revealed an aberrant overexpression of this gene in the second relapse sample of ALL1 (**Figure 3**). *FLT3* was highly expressed in the second relapse in both the absolute sense (i.e. when compared to all genes expressed in that tumor, **Figure 3C**), as well as in the relative sense (i.e. when compared to other B-ALL samples, **Figure 3B, 3D**). In the second relapse sample, *FLT3* had an estimated expression level (FPKM of 108.0) that placed it within the top 0.51% of all genes (**Table S5**). *FLT3* expression in the second relapse was also an outlier compared to blood cell types sorted from healthy donors, including hematopoietic stem and progenitor-enriched CD34+ fractions (**Figure 3B, 3C**) as well as additional B-ALL samples from Kang et al. 2010 (**Figure 3D**) [30]. The overexpression of *FLT3* was confirmed orthogonally by evaluation on an exon array platform, in triplicate, and was compared directly to exon array data from other studies (**Methods**).

We investigated whether cis-acting regulatory mutations might be increasing *FLT3* transcription levels. We scanned for potential regulatory somatic variants in the ALL1 second relapse WGS data. While we did not find any promoter mutations, we did find two somatic indels in a small region of *FLT3* intron 1 (**Figure S40**). Amplicon sequencing designed to flank both of these mutations revealed that they were not in linkage (i.e. both alleles had a different somatic *FLT3* intron 1 mutation). This raised the possibility that each mutation independently increased expression of its own allele. To test this, we made use of a heterozygous C/T single nucleotide polymorphism (SNP) in exon 2 (**Methods**). We reasoned that acquisition of each indel should result in an allele-specific increase in *FLT3* expression that could be assessed by the C:T ratio of the exon 2 SNP in RNA derived from that sample. Since only one of the indels was present at a high VAF at first relapse (day 1,893), mRNA isolated from that time point should predominantly represent a single allele, and should be associated with skewed expression of the SNP. RNA was purified from archived bone marrow biopsies taken at diagnosis (day 0), first relapse (day 1,893), and second relapse (day 3,072). As a control, RNA was isolated from a human B-ALL cell line (REH) that over-expresses *FLT3*, and is wild type for the exon 2 SNP (C/C). We used RT-PCR to amplify a short fragment containing the exon 2 SNP followed by digital PCR to compare the contribution of each allele to the total *FLT3* mRNA at each time point. We found that at each time point—diagnosis, first relapse, and second relapse—the relative contribution of each allele to total *FLT3* mRNA was equivalent, suggesting that the intron 1 indels do not promote *FLT3* transcription. Further, this SNP was expressed equally from both alleles in the second relapse sample in RNA-seq data (50.5% VAF based on 3,913 reads covering the SNP position). Therefore, these data clearly indicate that both *FLT3* alleles were massively overexpressed in the second relapse sample, suggesting that the gene was activated in trans by an unknown mechanism.

SUPPLEMENTARY FIGURES

Figure S1. Cytogenetics summary. (A) Results of a FISH using a break apart *ETV6* probe with 5' and 3' in red and green colors. A normal signal would be yellow and if it rearranged you will see a red and a green. In our case we just see a loss of one yellow- indicating that the whole *ETV6* locus was deleted. (B) A metaphase spread obtained at second relapse. Note evidence for a large scale deletion on chromosome 12 (where *ETV6* is located). (C) FISH for *ETV6* and *ETV6-RUNX1* fusion performed on tumor cells from the second relapse. Both assays indicate loss of a single copy of *ETV6* but the *ETV6-RUNX1* fusion is not detected.



A. *ETV6* (*TEL*) FISH of primary tumor nuclei

B. Metaphase spread obtained at second relapse

C. *ETV6* and *ETV6* (*TEL*) - *RUNX1* (*AML1*) FISH of second relapse nuclei

Figure S2. Sorting of blasts and lymphocytes from the second relapse
The second relapse sample used for WGS and variant discovery was obtained at day 3,072 after the patient had received an allograft from his brother. To obtain a sample from the second relapse that would not be substantially contaminated by cells from the sibling donor, the bone marrow sample obtained at second relapse was subjected to the following sort. Blasts and lymphocytes were sorted concurrently from the same second relapse sample by CD45 and side scatter (lymphocytes were CD45 bright with low side scatter and blasts were CD45 dim with low side scatter). We then collected CD19+/CD34+ cells in the blast gate. This sample is referred to as 'SB_d3072_A' in this work. As a control, the lymphocytes were also sorted and subjected to custom capture sequencing along with samples from other time points throughout disease progression (**Methods**). This sample is referred to as 'SL_d3072_I'.

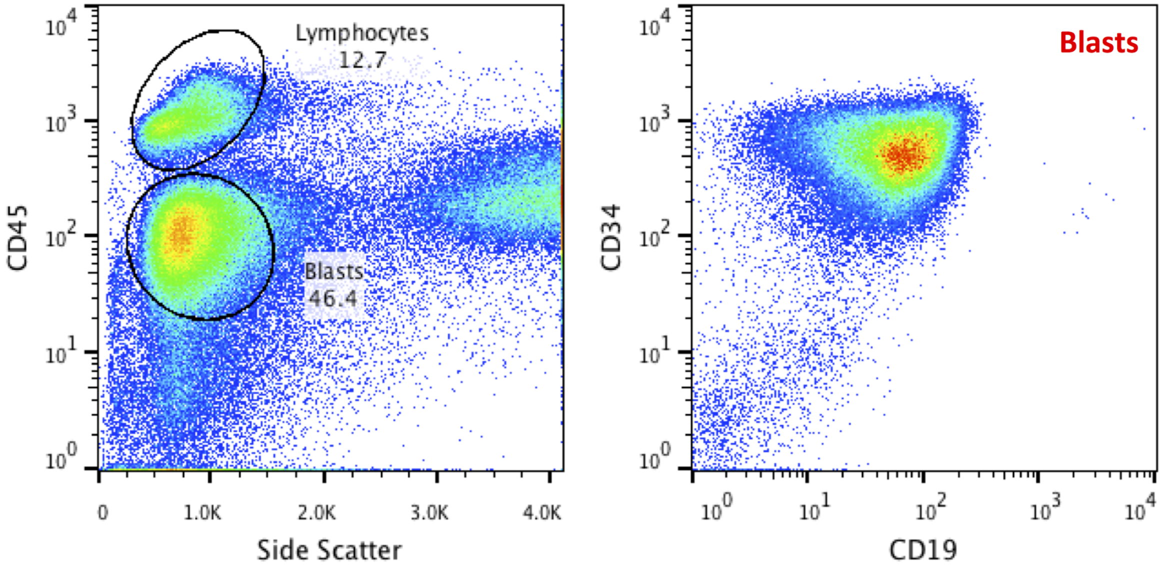


Figure S3. Sequencing coverage summary

Coverage depths for sequencing alignments of samples in this study are summarized as heatmaps. Data are shown for whole genome sequencing (A), custom capture sequencing (B), exome capture sequencing (C), and RNA sequencing (D).

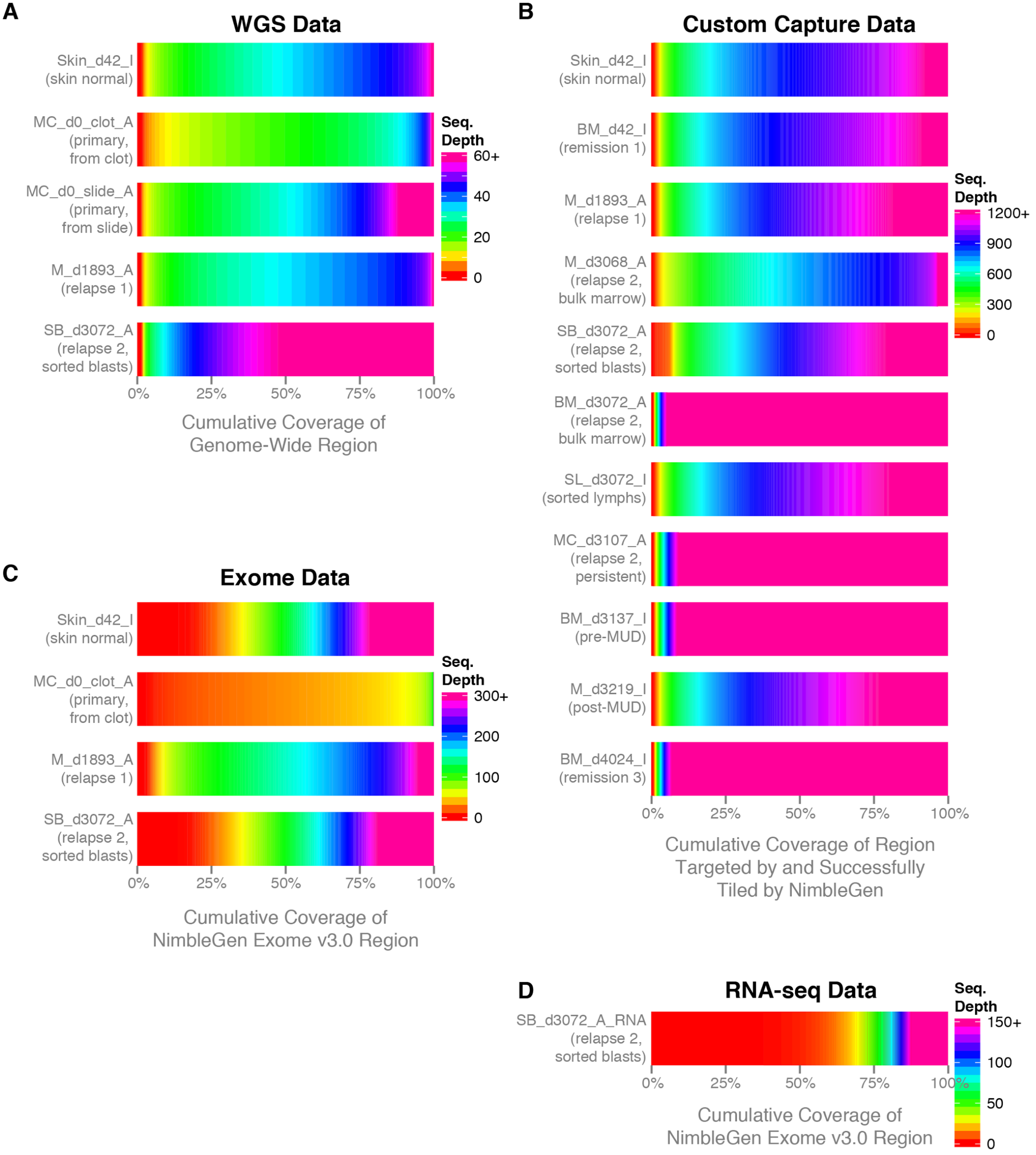


Figure S4. Identity SNP analysis of all sample subjected to whole genome sequencing
A summary of 24 ‘identity SNPs’ used to confirm the genotype of five samples obtained from our patient and subjected to whole genome sequencing. Variant allele frequencies (VAFs) were calculated for these 24 SNPs for all five samples with WGS data. (A) Each position along the x-axis indicates a SNP position with reference base indicated at the bottom and variant base indicated at the top. VAFs are plotted on the y-axis. If a sample is heterozygous for a SNP the VAF should be around 50%. If a sample is homozygous the VAF should be close to 0% or 100%. All samples appear to agree with respect to their genotypes for these identity SNPs suggesting that this data are not affected by sample swaps or contamination from unrelated DNA. (B) The coverage or total read count observed at each position in the WGS data for each sample is shown. All samples achieved at least 10x (generally 30-50x) coverage at the 24 identity SNP positions.

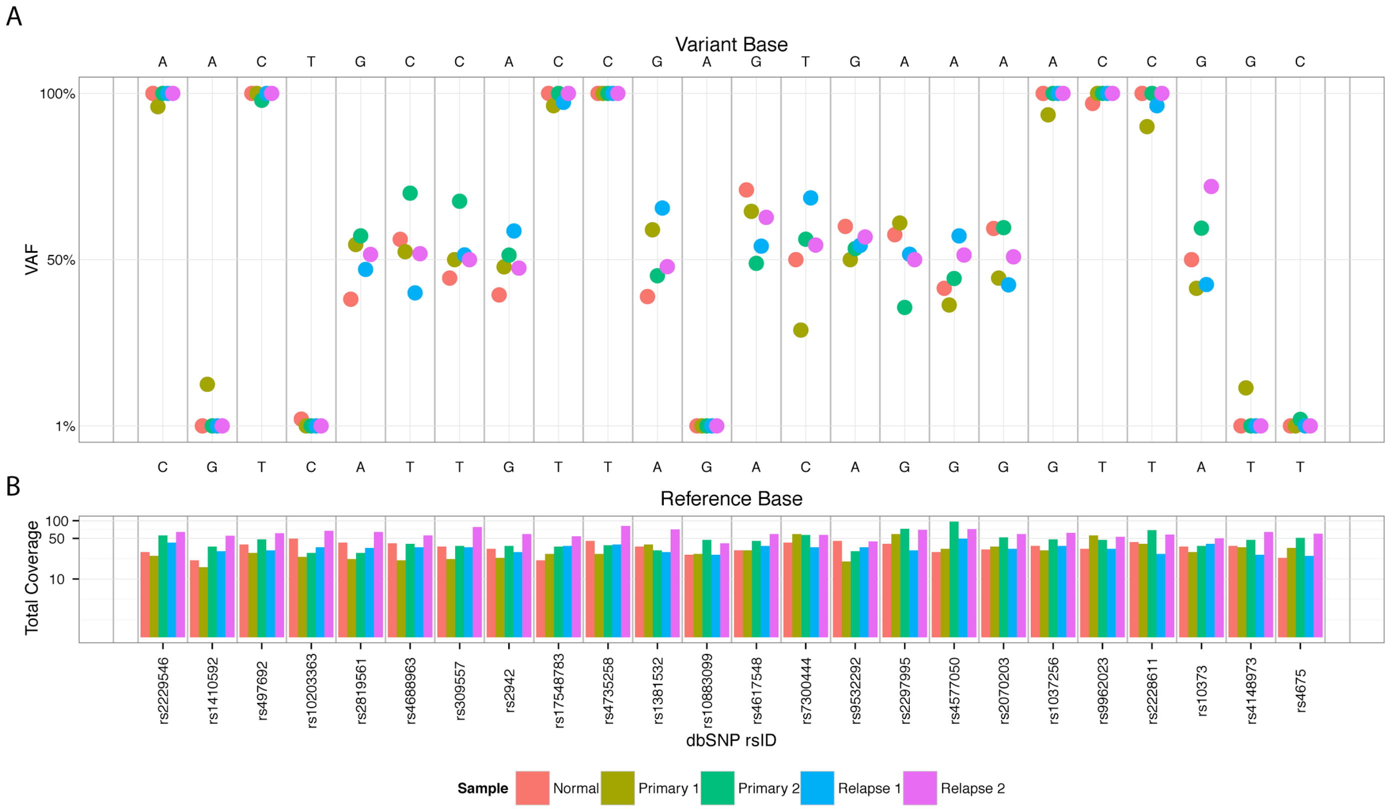


Figure S5. Mutation spectrum (Tv/Ti) analysis
Transitions/Transversions present in the master variant list (**Table S7**) were calculated and plotted via the R package GenVisR for the second relapse sample (SB_d3072_A). Coverage filters were applied requiring 50x depth in both tumor and normal samples. Variant allele frequency (VAF) filters requiring >= 20% in tumor and <= 1% in normal were also applied. SNV locations passing filters were randomly mutated with equal probability (⅓) 100 times and the proportions for each transition/transversion were calculated. The average of these calculations are displayed as expected values (left). The average variance of this calculation was found to be 6.02e-5.

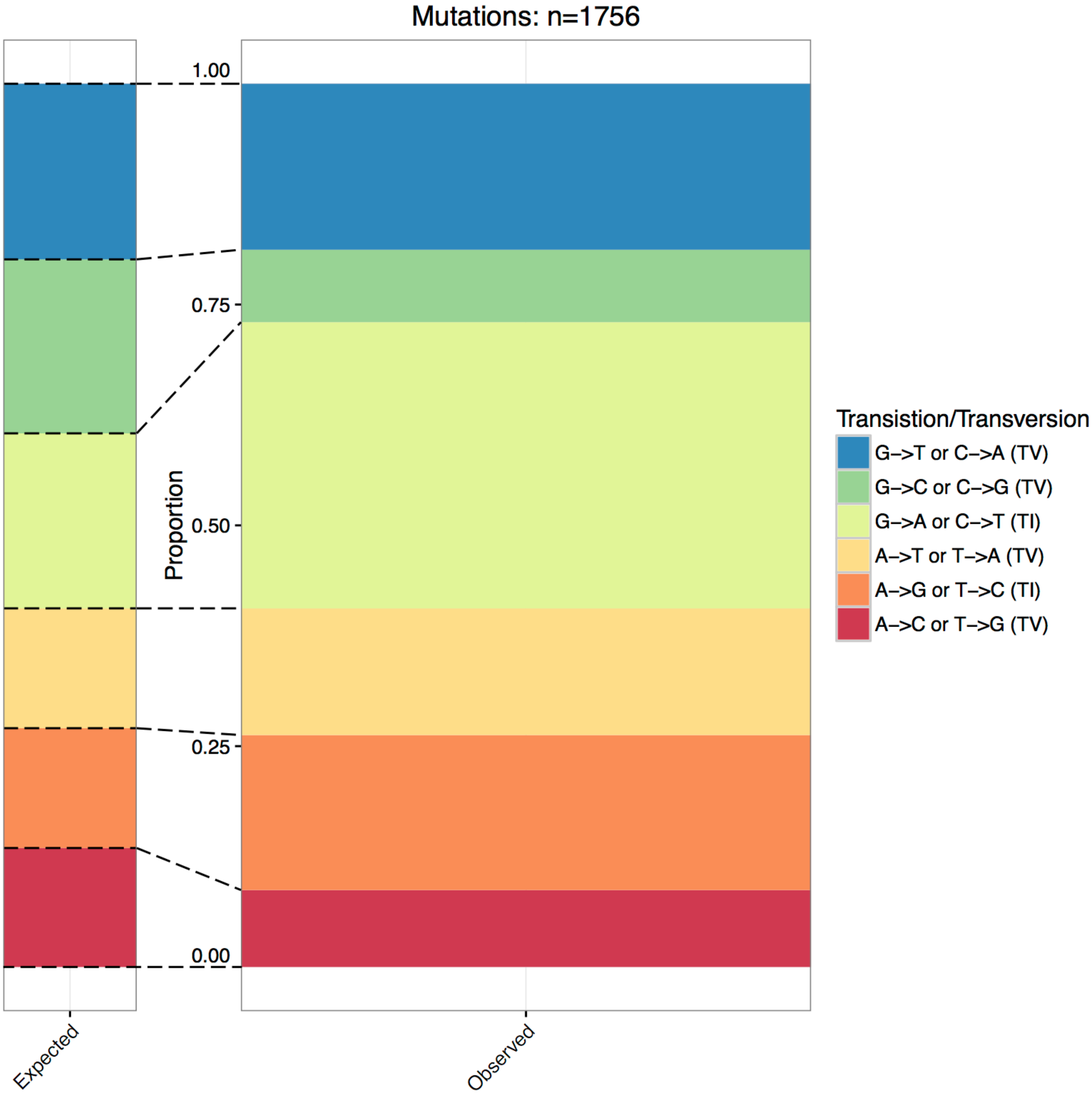


Figure S6. Estimation of allograft donor DNA contamination

Sequencing and variant detection was performed on some samples obtained after a sibling allograft (on day 2,010). To determine the possible contribution of this DNA to false positive somatic variants called in the second relapse, we estimated the proportion of contaminating DNA present from the sibling in each sample (**Methods**, **Table S1**). In the top panel, VAFs for heterozygous germline SNPs called from the normal skin are plotted for each time point. Heterozygous variants in the normal (group A) are highlighted in red. Variants that were homozygous after the first allograft in the sorted lymph (group C) are highlighted in blue. These variants were used to estimate contamination from the sibling donor (**Table S1**). In the second panel the VAFs for this subset of variants are shown. As expected contamination is not detected prior to the allograft. The sorted blasts sample obtained at second relapse (day 3,072) also shows very little contamination indicating that the sorting strategy (**Methods**) was successful in purifying tumor cells from the patient. Bulk marrow samples obtained from the same time point show considerable levels of contamination from the sibling donor. In the third panel, heterozygous variants in the normal (group A) that were also heterozygous in the sorted lymph (group B, highlighted green) but homozygous after the second allograft (group D, highlighted purple) were selected to estimate contamination from the second allograft. Note that only the following samples, days 3,219 and 4,024 were obtained after the second MUD allograft (on day 3,151) and therefore contamination from this genotype is only expected in these samples. Vertical lines indicate time points for bone marrow transplants. The fourth panel refers to the tissue type and sorting strategy (if any) applied to each sample.

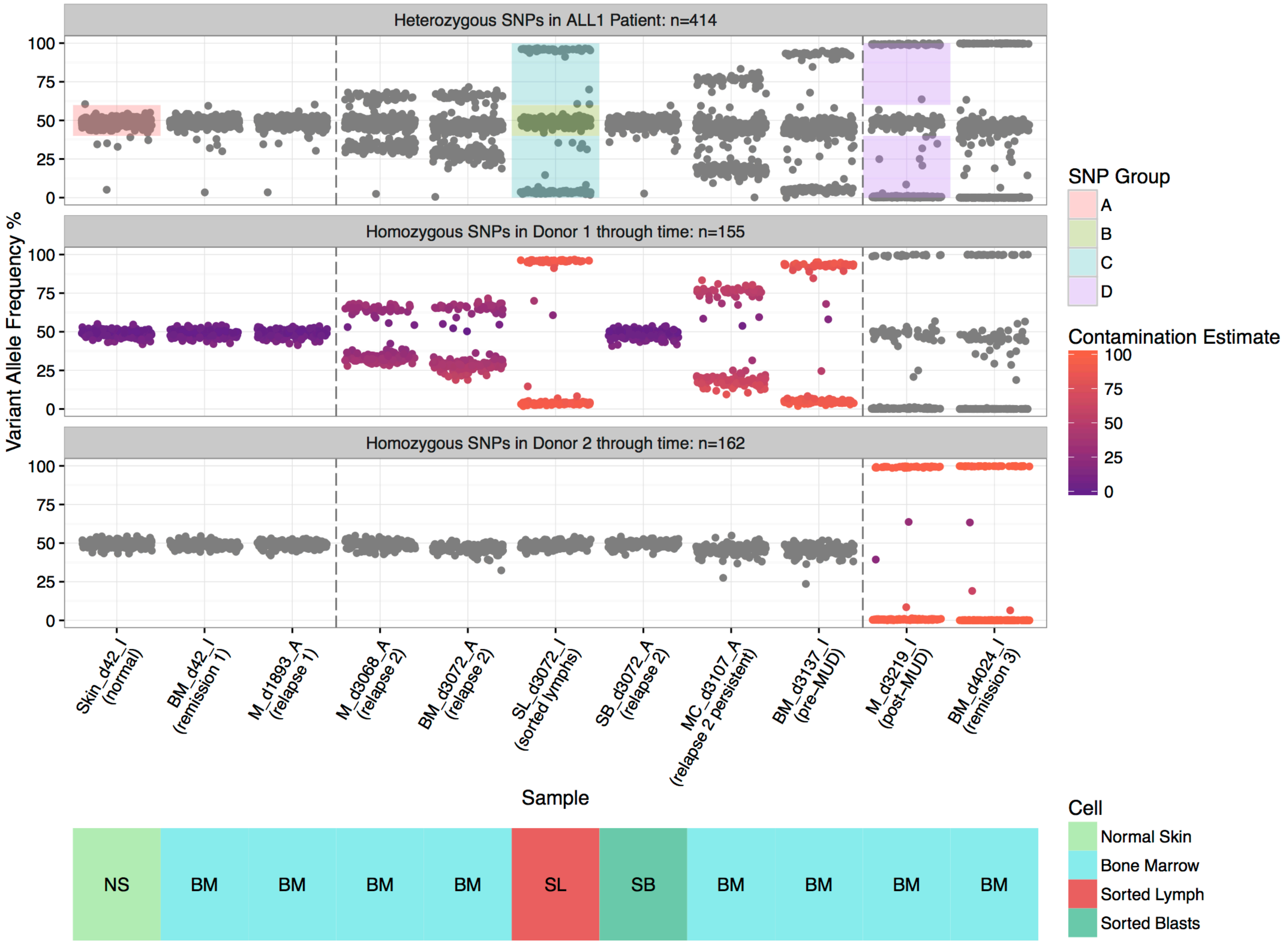
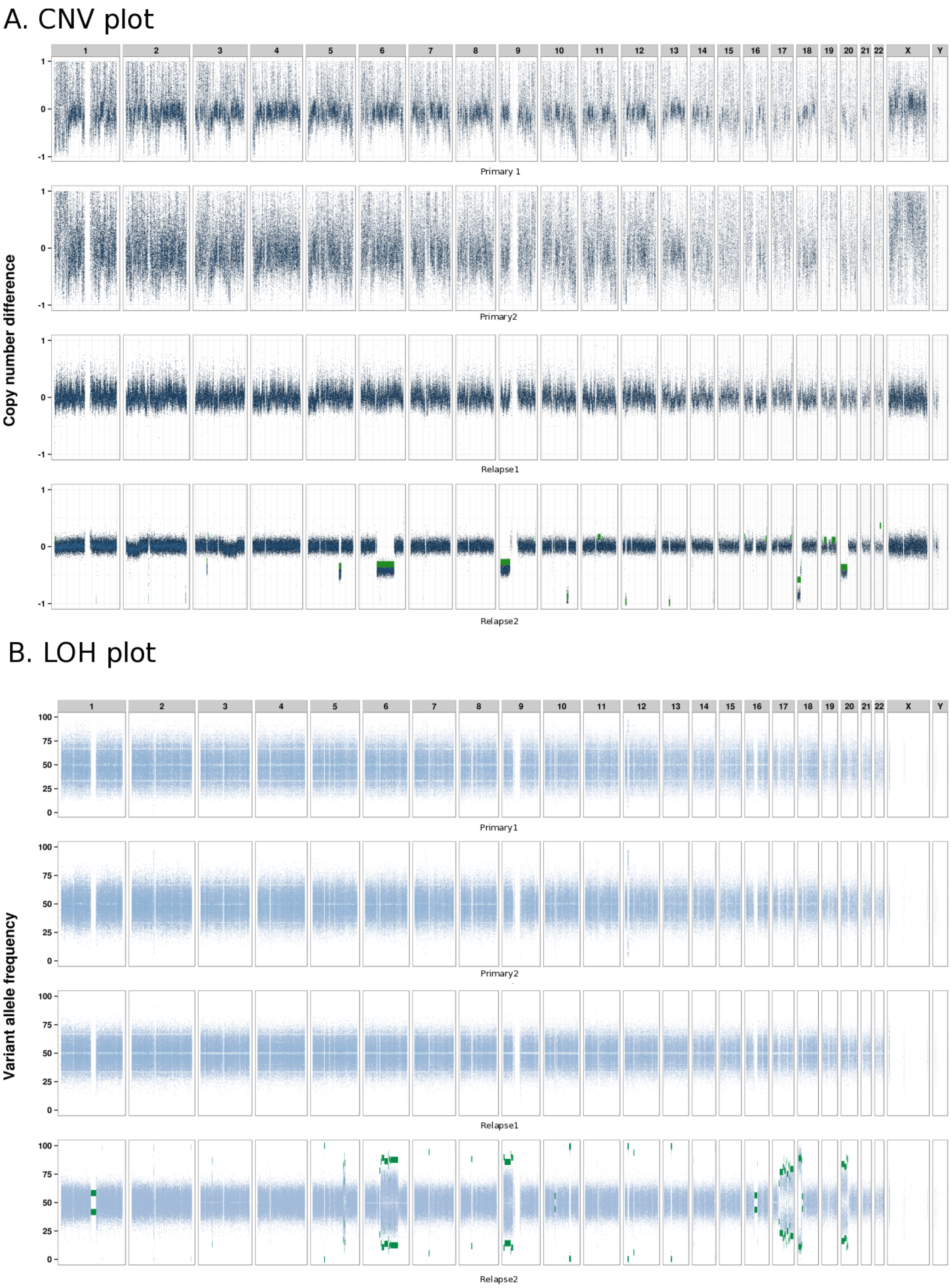
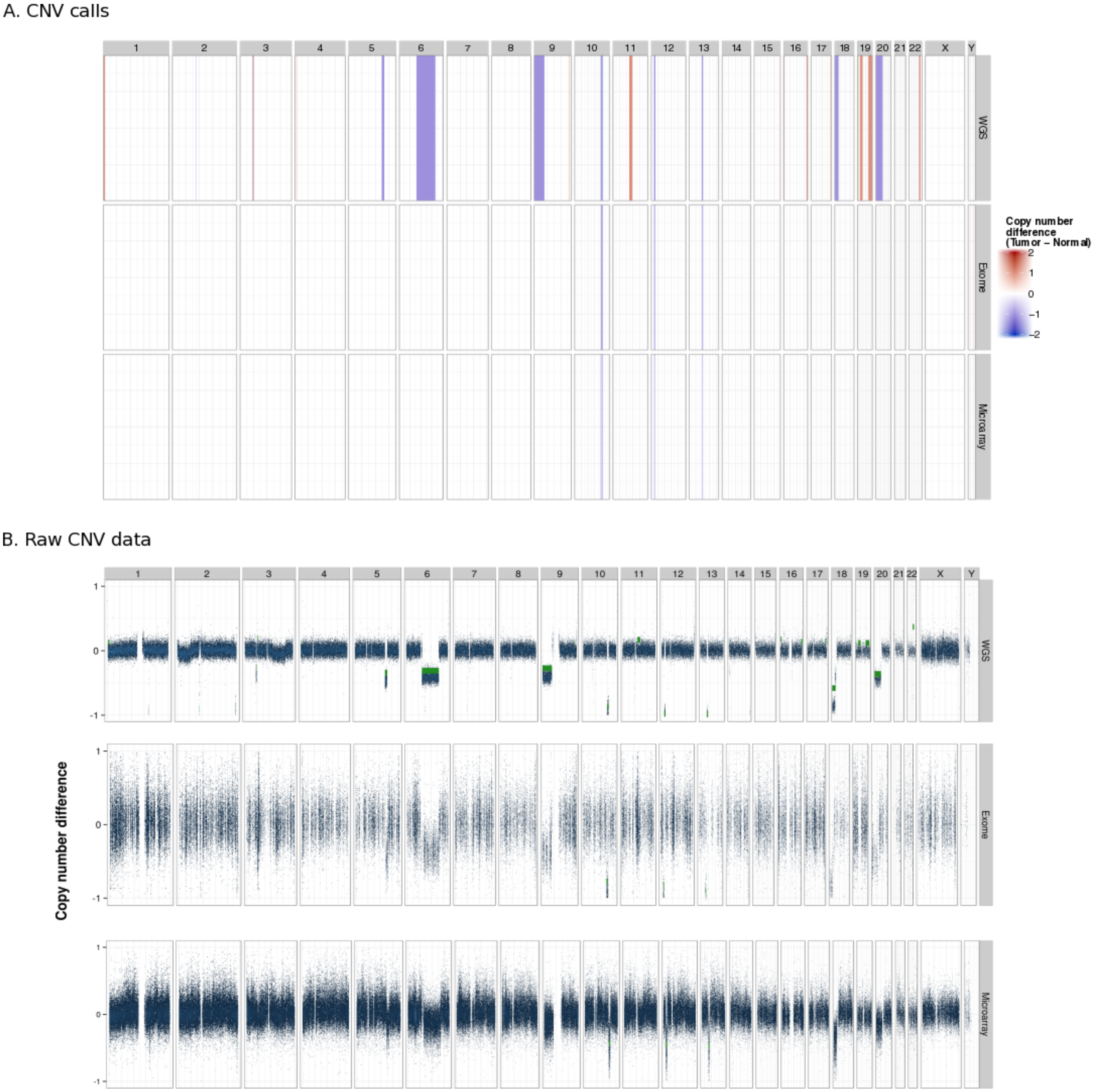


Figure S7. Genome wide CNV and LOH analysis of whole genome data
(A) Copy number differences (tumor - normal) for 10 kb coverage windows across the entire genome are plotted against chromosome position for all four tumor samples with whole genome sequence data (one blue dot for every 10 kb window) (**Methods**). Due to low amounts of input DNA or degraded material for both primary samples, and low purity of the first relapse, confident calling of copy number variation (CNV) events was not possible in these samples. The high quality and abundant material obtained from the second relapse allowed the identification of CNV segments indicated in green (B) Loss of heterozygosity was assessed by examining variant allele frequencies (VAFs) for variants found to be heterozygous in the normal skin sample.


A. CNV plot

B. LOH plot

Figure S8. Copy number calls across platforms for the second relapse
Copy number variation was identified using WGS, exome and microarray data. (A) Segmented calls identified from the different platforms are displayed for all chromosomes. CopyCat (https://github.com/chrisamiller/copycat) was used to identify CNVs from the WGS data, CnMops [18] was used to identify CNVs from exome data and DNACopy [57] was used to identify CNVs from microarray data. (B) The raw data that was used to identify the CNVs in the previous panel plotted against chromosome position for all chromosomes. The raw data consists of normalized read-depth values for the WGS data (one for each fixed 10kb window across the genome), normalized read-depth values for the exome data (one for each exon targeted by the exome reagent), and probe intensities for the microarray data.


A. CNV calls

B. Raw CNV data

Figure S9. Analysis of clonal architecture of the second relapse (sorted blasts) sample
The clonal architecture of the second relapse sorted blasts tumor sample was assessed by examining the distribution of variant allele frequencies (VAFs) for 1,588 high quality somatic variants. High quality somatic variants were those detected in the WGS data and validated by deep capture sequencing. VAFs used in this analysis were calculated using all combined data (WGS, exome, and custom capture). Variants in regions of CNV or LOH were also removed because of the increased uncertainty for VAFs of these variants (**Methods**). (A) VAFs were clustered to allow identification of potential subclones and plotted as a density plot. These clusters were labeled as the 'Founding' clone (green), 'Subclone A' (purple), and 'Subclone B' (blue). 'Founding' clone variants are those that appear to be in every cell of the second relapse tumor sample while 'Subclone A' and 'Subclone B' variants represent variants present in only a subset of cells of the tumor. (B) Coverage values for the same high quality somatic variants were plotted against VAF and colored according to subclone assignment. Coverage values were ceilinged at 2,000x for display purposes.

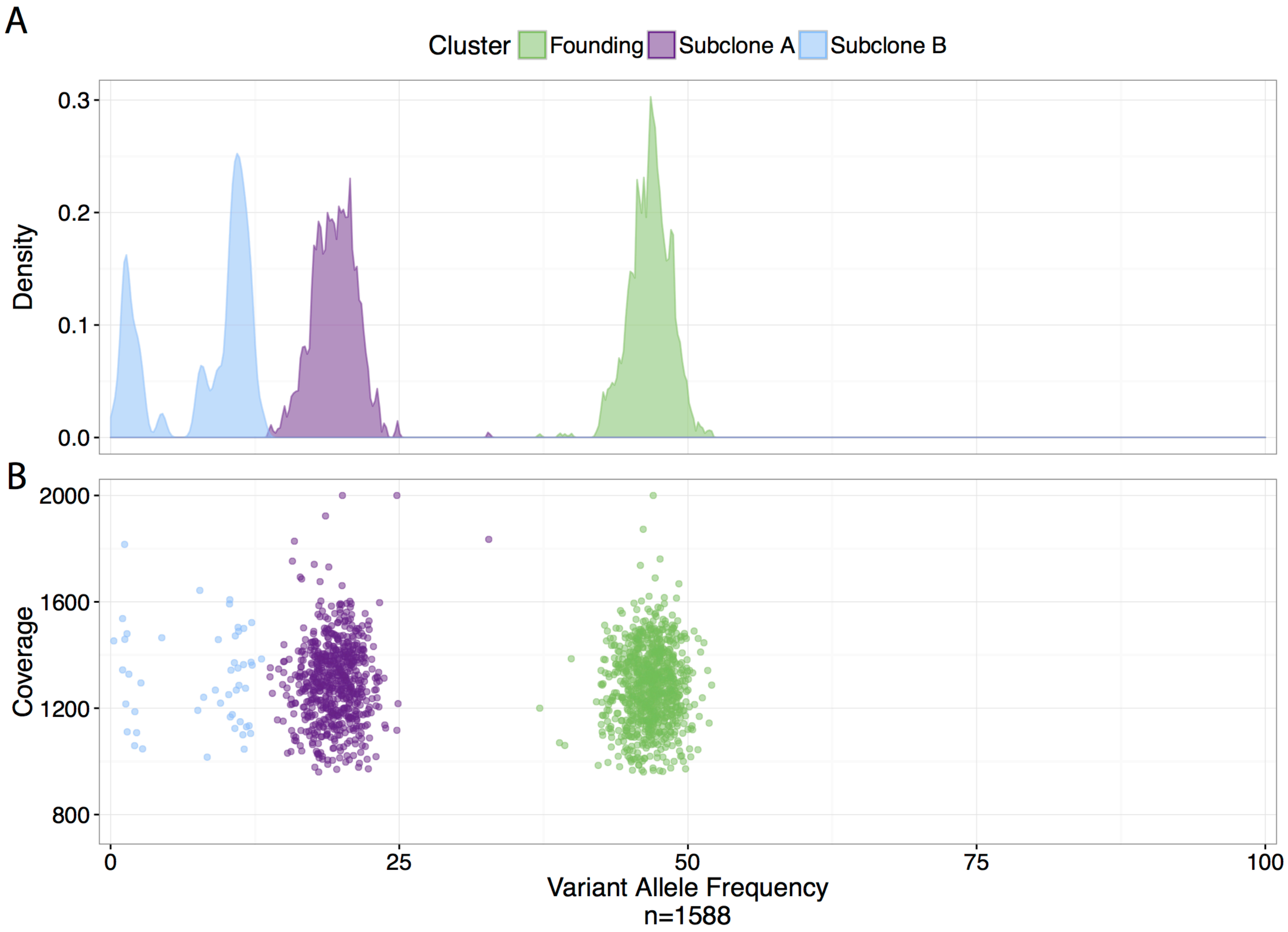
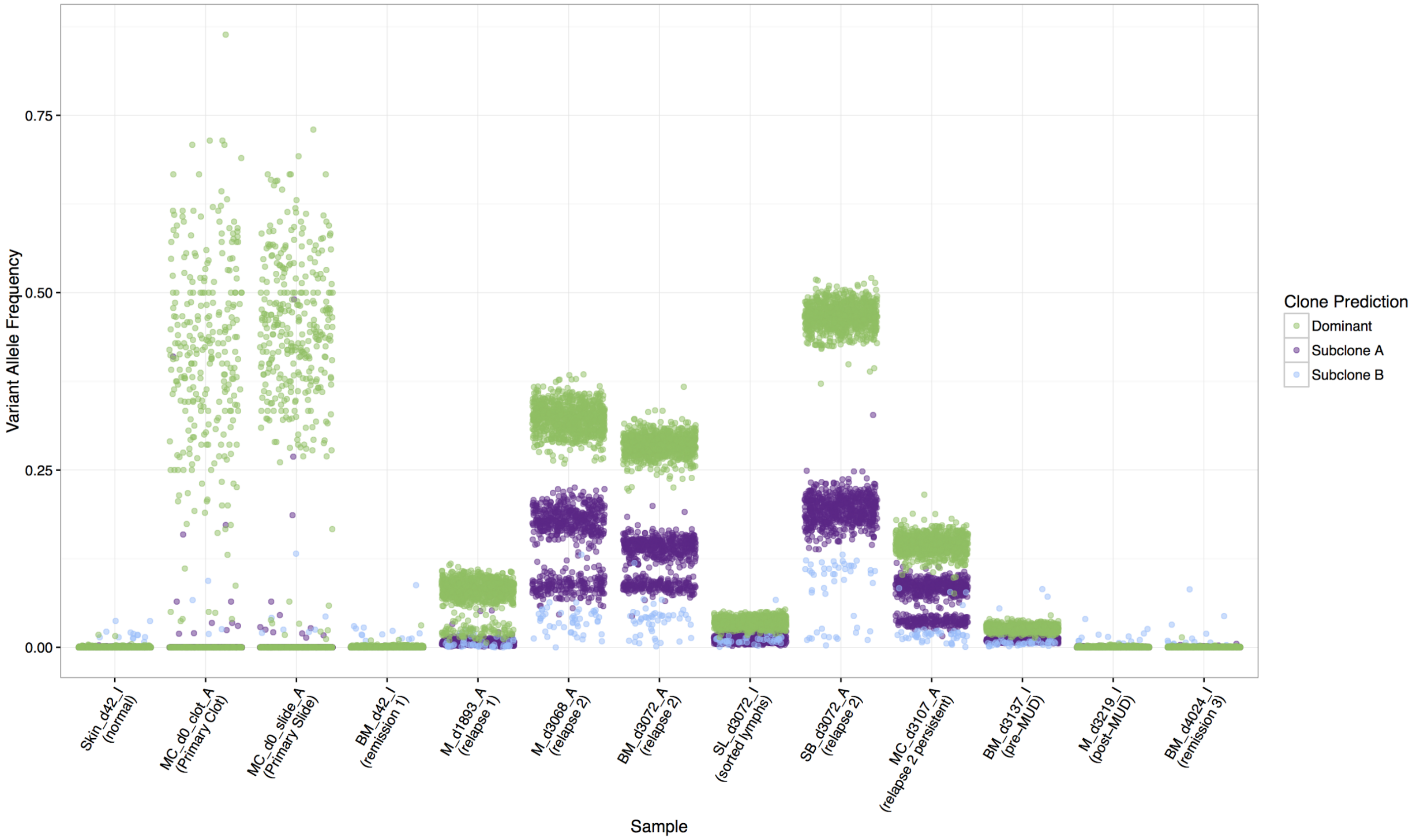


Figure S10. Estimation of tumor purity at all time points
Display of capture variant allele frequencies (VAF) originating from the relapse 2 sorted blast sample (SB_d3072_A) across 12 additional samples. Clonal predictions are derived from relapse 2 (SB_d3072_A) and are used to colour variants across all time points. A coverage filter of 20x was applied to all variants at each individual time point to ensure accurate VAF estimates. Estimated tumor purity and the number of variants passing filters at each time point are displayed in the lower panel.



Cell Type	NS	BM	BM	BM	BM	BM	BM	SL	SB	BM	BM	BM	BM
% Purity	0.1976	83.87	89.27	0.2031	16.21	64.72	57.05	7.05	93.6	28.94	5.403	0.1803	0.07994
Variants	1588	1311	1532	1588	1588	1588	1588	1588	1588	1588	1588	1587	1587

Cell Type
Normal Skin
Bone Marrow
Sorted Lymph
Sorted Blast

Figure S11. Pairwise comparisons of VAFs for all coding variants in primary, relapse 1 and relapse 2
Variant allele frequencies (VAFs) for 98 variants predicted to affect coding sequences (**Table S3**) were plotted for all pairwise comparisons of the two primary tumor samples, relapse 1 and relapse 2. Where applicable, VAFs were calculated from combined WGS, exome, and capture data for each sample (**Table S8**). Inset panels are used when VAFs are low due to low tumor purity. Black dots indicate selected high VAF variants labeled with a gene symbol. Note variants along each axis that represent variants that were potentially gained or enriched during tumor evolution. For example, an *NF1* variant was present in relapse 1 and relapse 2 but was undetectable in either primary sample. By contrast two *EP300* variants were detectable in both primary samples as well as relapse 1 and 2.

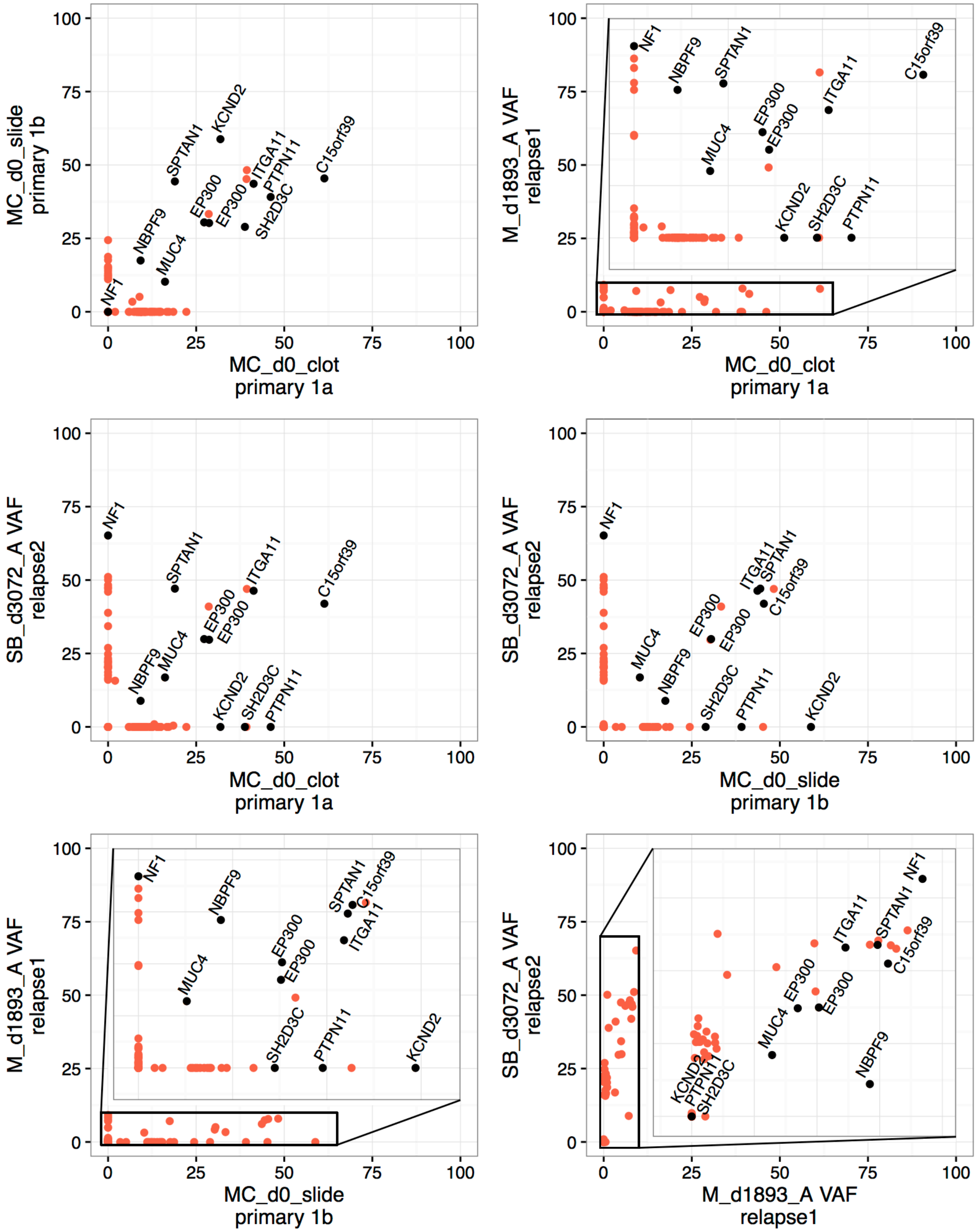


Figure S12. Detection of coding variants throughout disease progression
Select coding somatic variants originating from the WGS datasets and passing manual review (**Table S3**) are displayed as bar plots of variant allele frequencies (**Table S8**) for multiple samples. To be included in this analysis, variants must have passed review in either relapse-1/relapse-2 alone or in any two samples. Situations in which coverage was less than 10 reads are annotated as red triangles. Clustering is based on VAF values derived from **Figure S9**, and variants that are copy altered, or are primary specific are indicated.

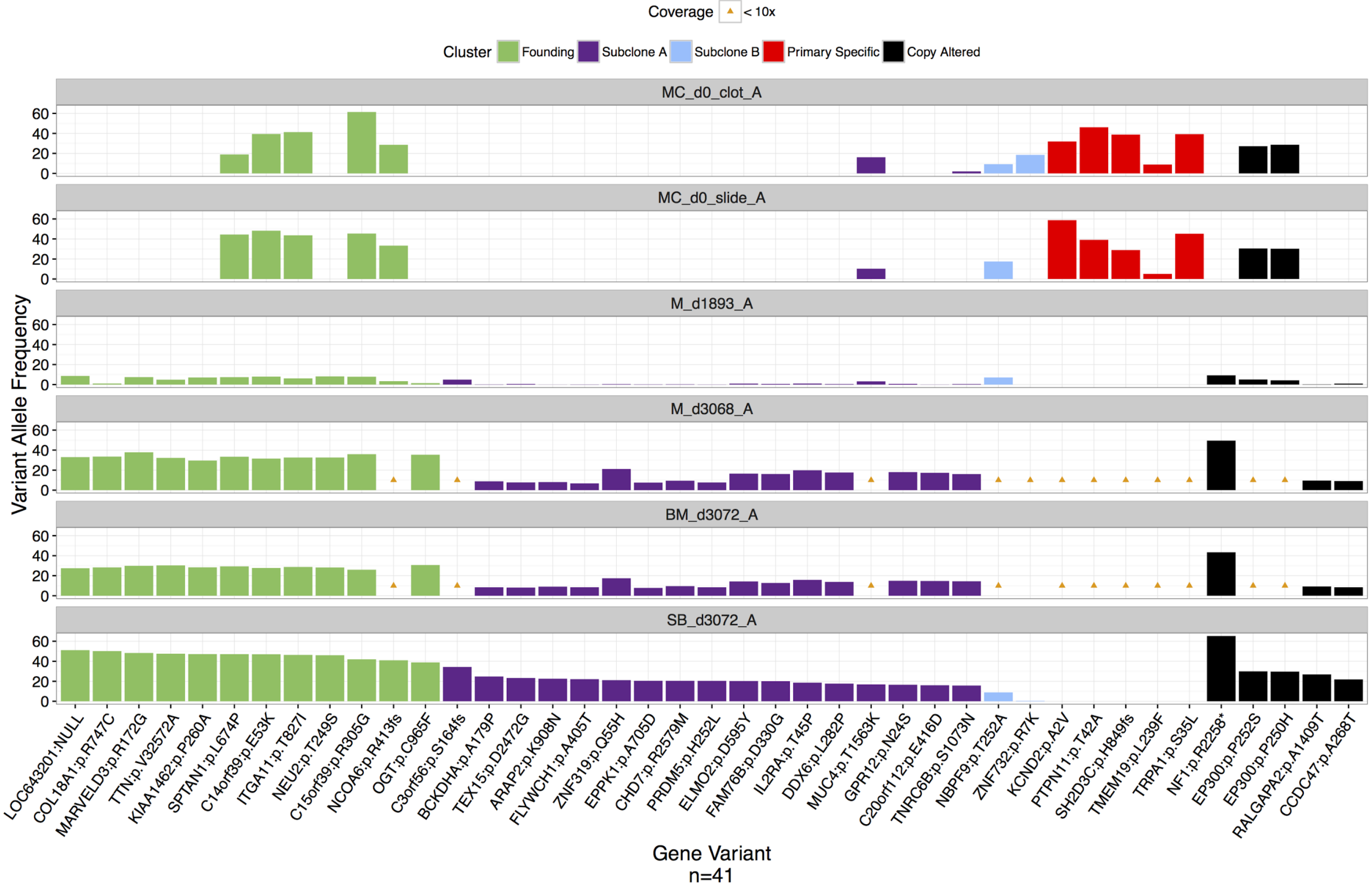


Figure S13. Read support for two linked missense SNVs in *EP300* (P250H and P252S)
(A) Readcounts for reference (green) and variant (red) alleles are displayed for two *EP300* variants (p.P250H, p.P252S) identified in the primary among 6 samples representing 4 time points. (B) Variant allele frequencies (VAFs) at these time points are displayed among the 6 samples. (C) Mutations identified in Cosmic are displayed opposite the *EP300* variants from this case in relation to known protein domains for the transcript ENST00000263253. The expression level (FPKM value) for *EP300* is 17.4.

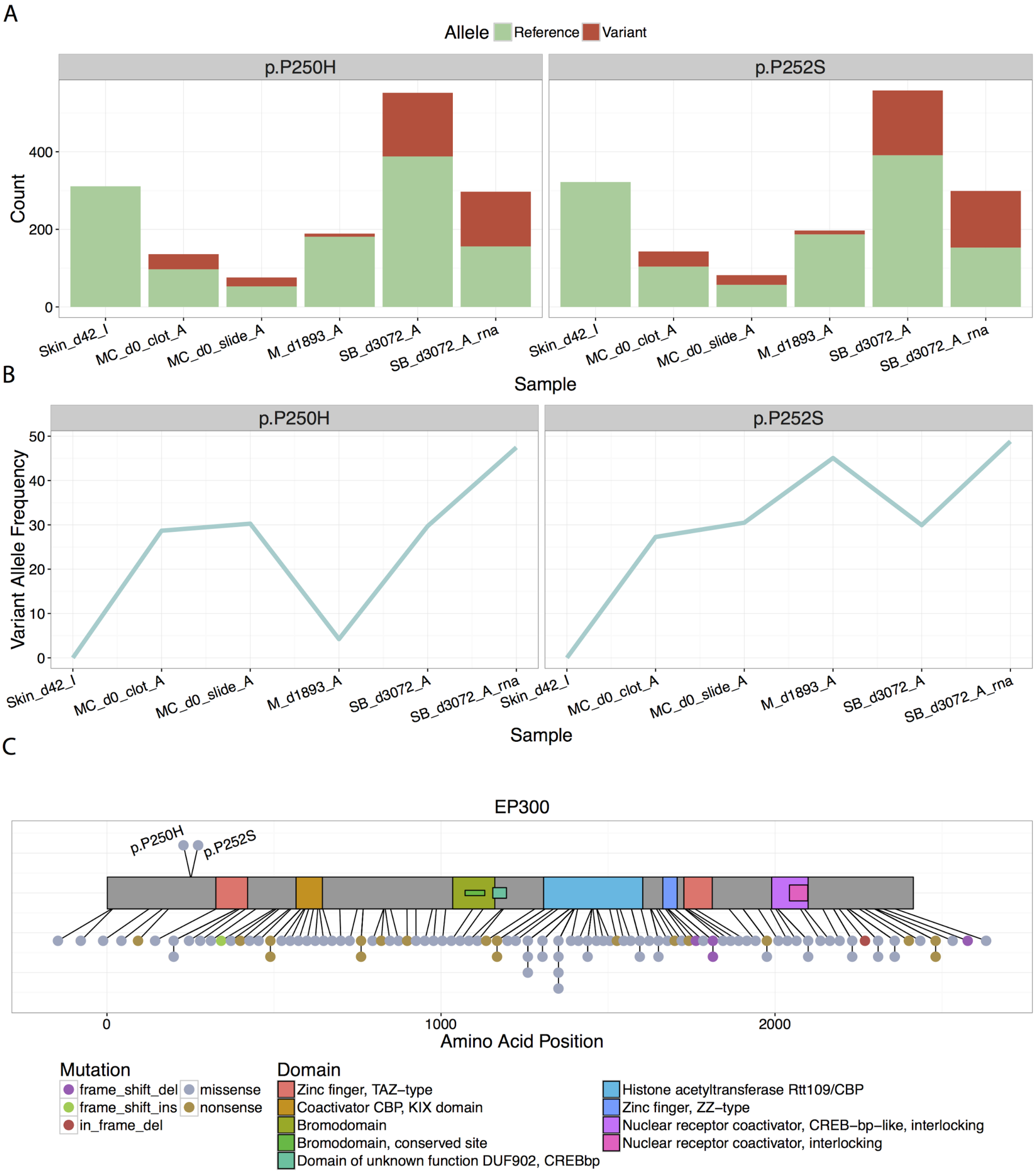


Figure S14. Read support for an *NF1* nonsense SNV (R2258*)

(A) Readcounts for reference (green) and variant (red) alleles are displayed for an *NF1* p.R2258* variant identified at the first relapse among 6 samples representing 4 time points. (B) Variant allele frequencies (VAFs) at these time points are displayed for the 6 samples. (C) Mutations identified in Cosmic are displayed opposite the *NF1* p.R2258* variant in relation to known protein domains for the transcript ENST00000358273. The expression level (FPKM value) for *NF1* is 12.2.

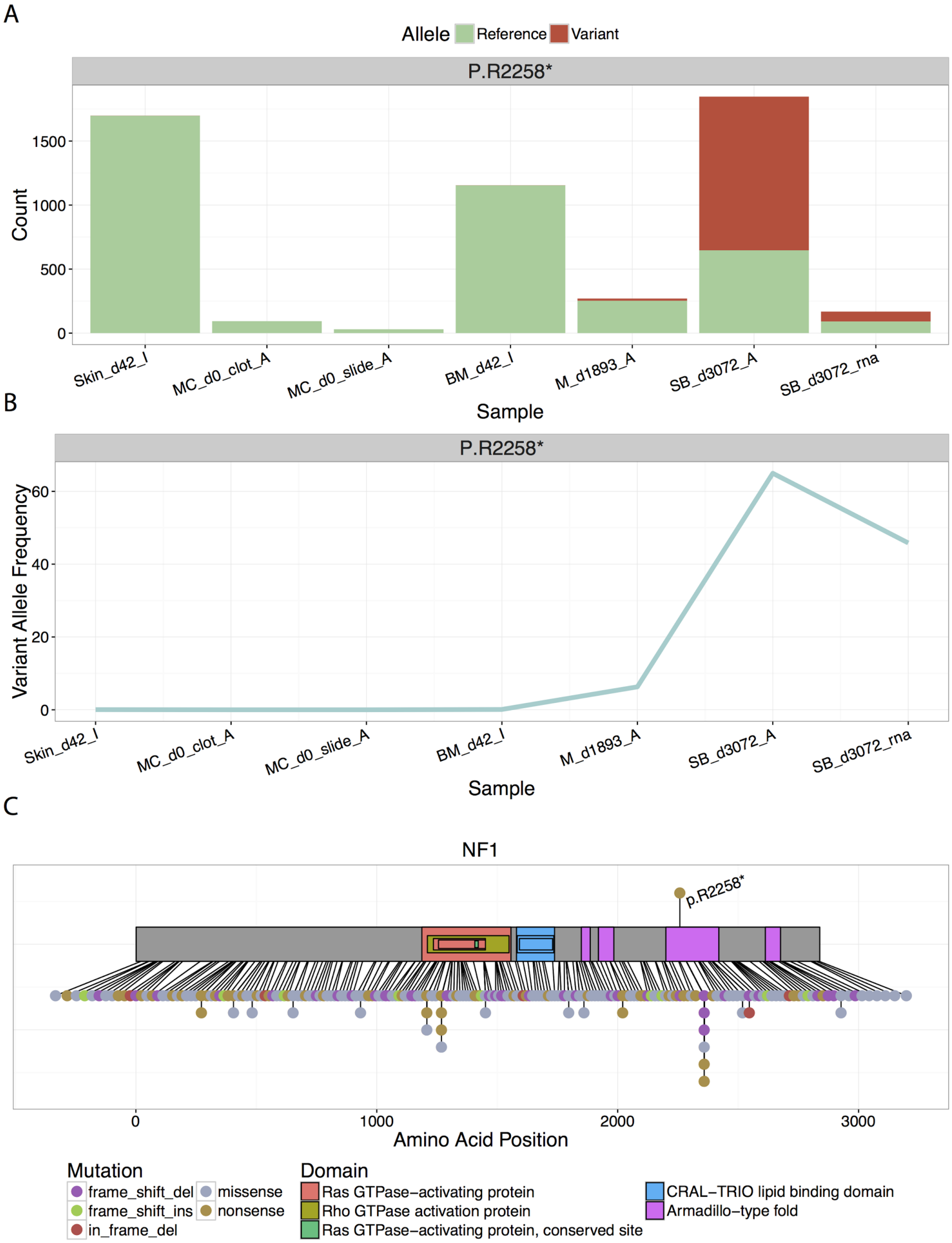


Figure S15. Read support for a *SETD2* frameshift indel (R2510fs)
(A) Readcounts for reference (green) and variant (red) alleles are displayed for a *SETD2* p.R2510fs variant identified at the second relapse among 6 samples representing 4 time points. (B) Variant allele frequencies (VAFs) at these time points are displayed among the 6 samples. A VAF is not shown for the RNA sample because indels of this size could not be detected by our RNA-seq alignments. (C) Mutations identified in Cosmic are contrasted with the ALL1 variant in relation to known domains for transcript ENST00000409792. The expression level (FPKM value) for *SETD2* is 17.6.

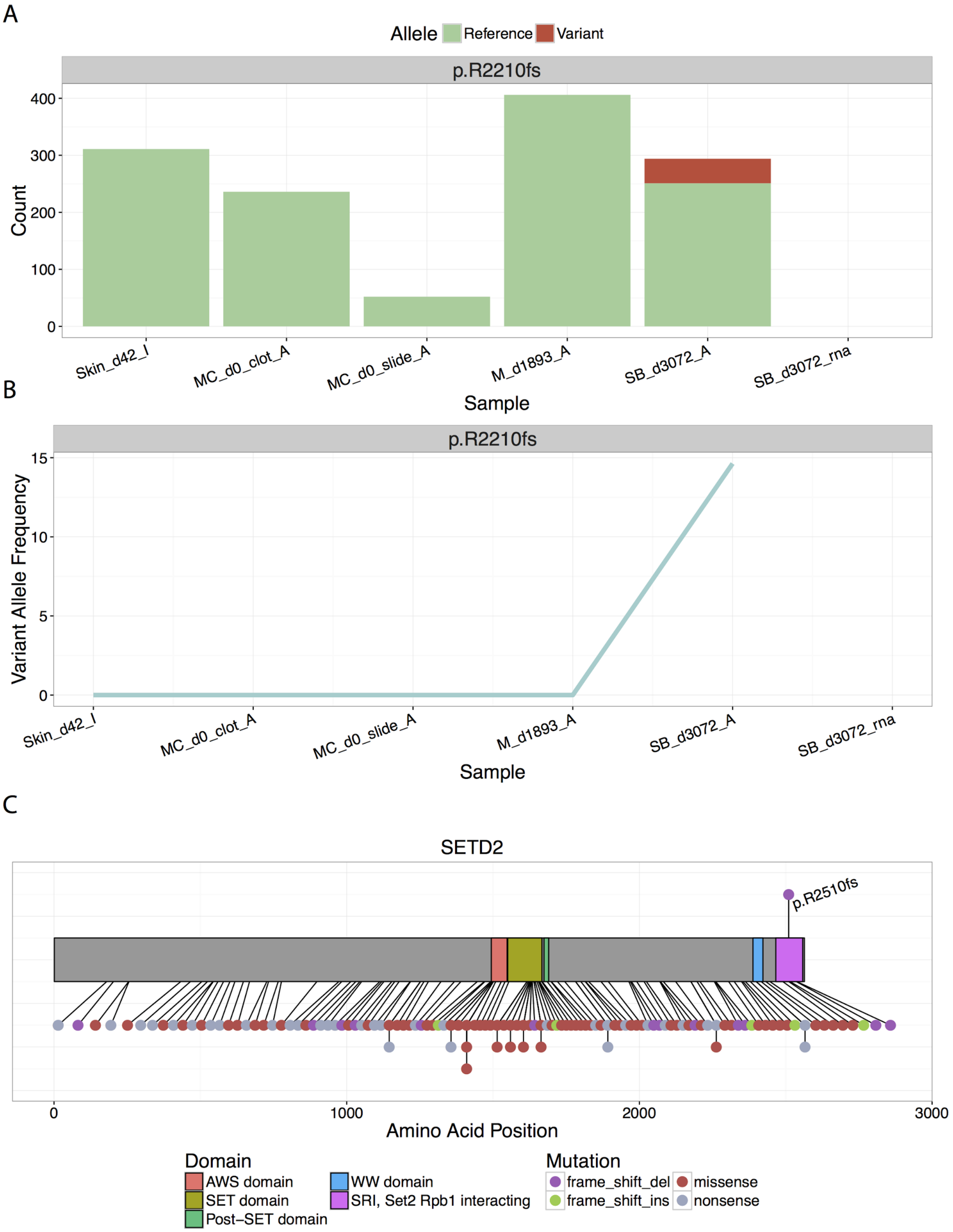


Figure S16. Focal amplification at 1p36.13 - 1p36.12 (near *SPEN* and *SDHB*)

(A) Ideogram showing Giemsa banding and labels for chromosome 1. (B) Copy number differences (CNV) (y-axis) were calculated as the difference between the second relapse tumor (SB_d3072_A) and skin normal and plotted against chromosome position (x-axis). Each black point represents the result of this calculation for WGS read coverage data from a single 10 kb window. Negative values indicate copy loss in the tumor and positive values indicate copy gain in the tumor. Boundaries of the region of interest (ROI) summarized in the legend are indicated as dotted vertical lines. Points colored red indicate copy number amplification/gain and blue points indicate copy number deletion/loss. (C) Variant allele frequencies (VAFs) in the tumor are plotted for all SNPs identified as heterozygous in the normal sample. Variants deviating from 50% (heterozygous A/B) towards 0% (homozygous A/A) or 100% (homozygous B/B) indicate a region of loss of heterozygosity (LOH). Points for SNPs within the ROI are indicated as green circles and SNPs outside the ROI are indicated as black circles. (D) CNV differences (red, blue, or black circles) with LOH VAF data overlaid (as green or grey triangles) are plotted for a magnified view of the ROI. Cancer gene positions are shown. Refer to **Methods** for more details.

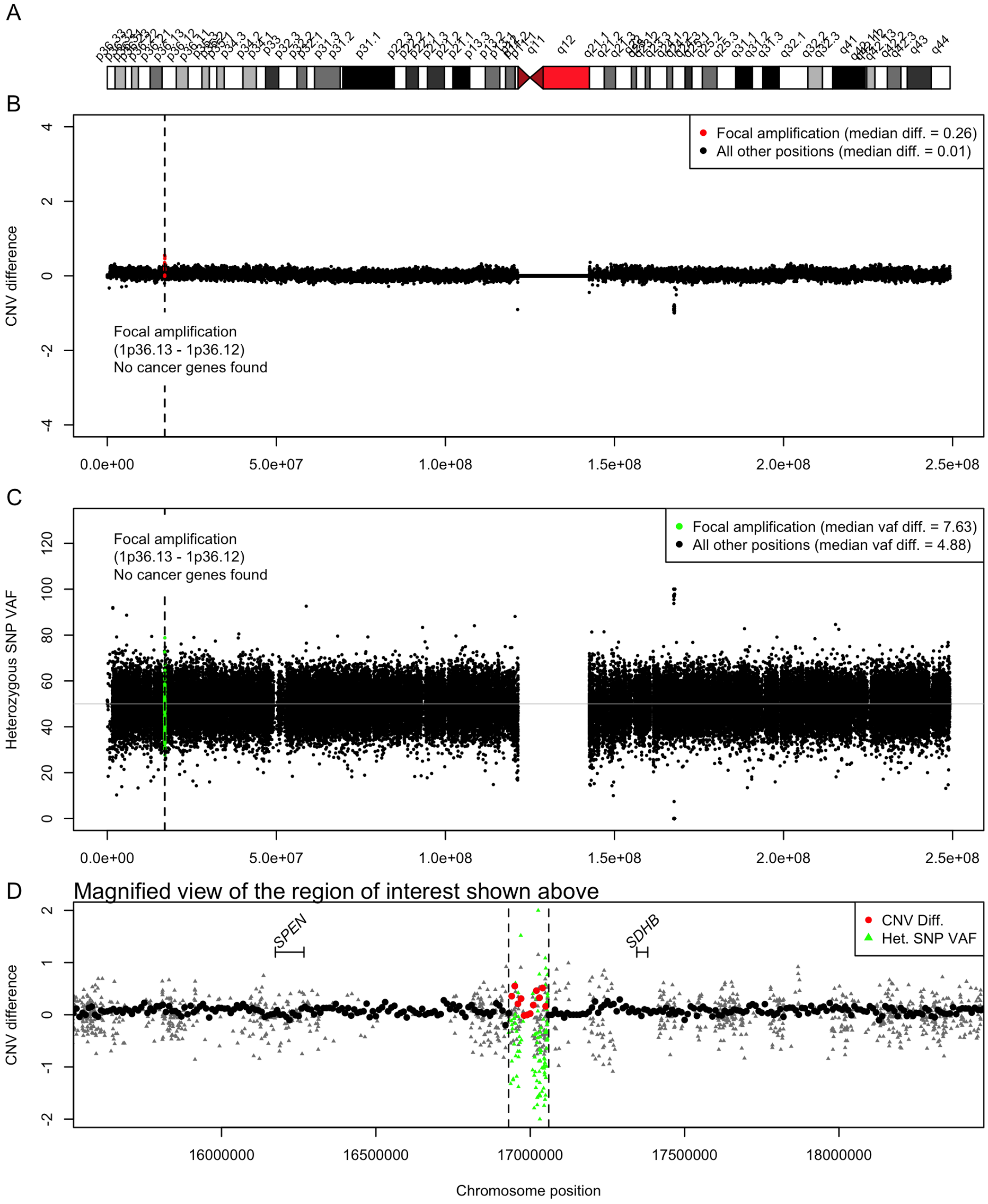


Figure S17. Focal deletion at 3p21.31 (affecting *SETD2*)

(A) Ideogram showing Giemsa banding and labels for chromosome 3. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

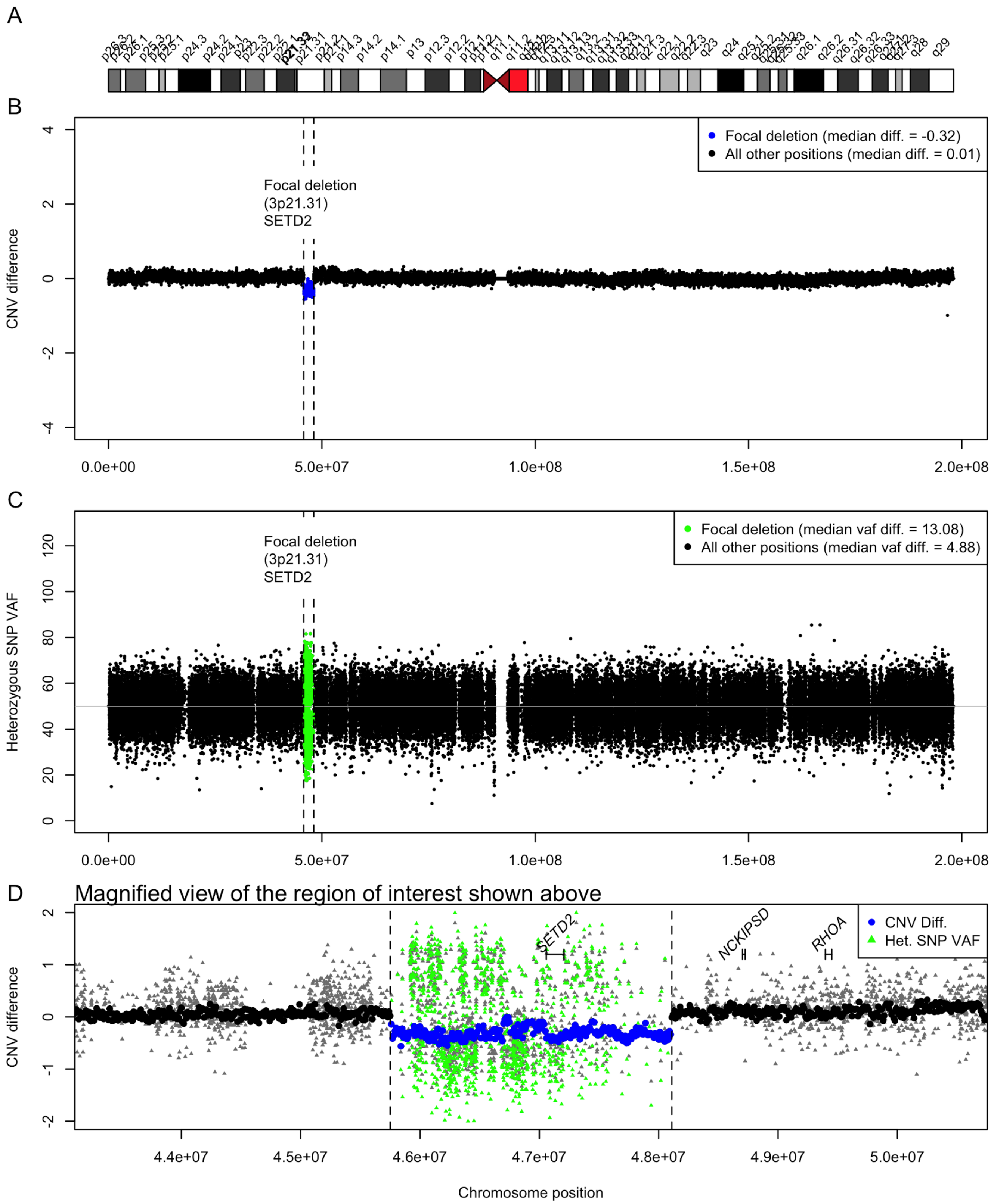


Figure S18. Large deletion at 5q23.2 - 5q31.1 (affecting *ACSL6* and *AFF4*)
(A) Ideogram showing Giemsa banding and labels for chromosome 3. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

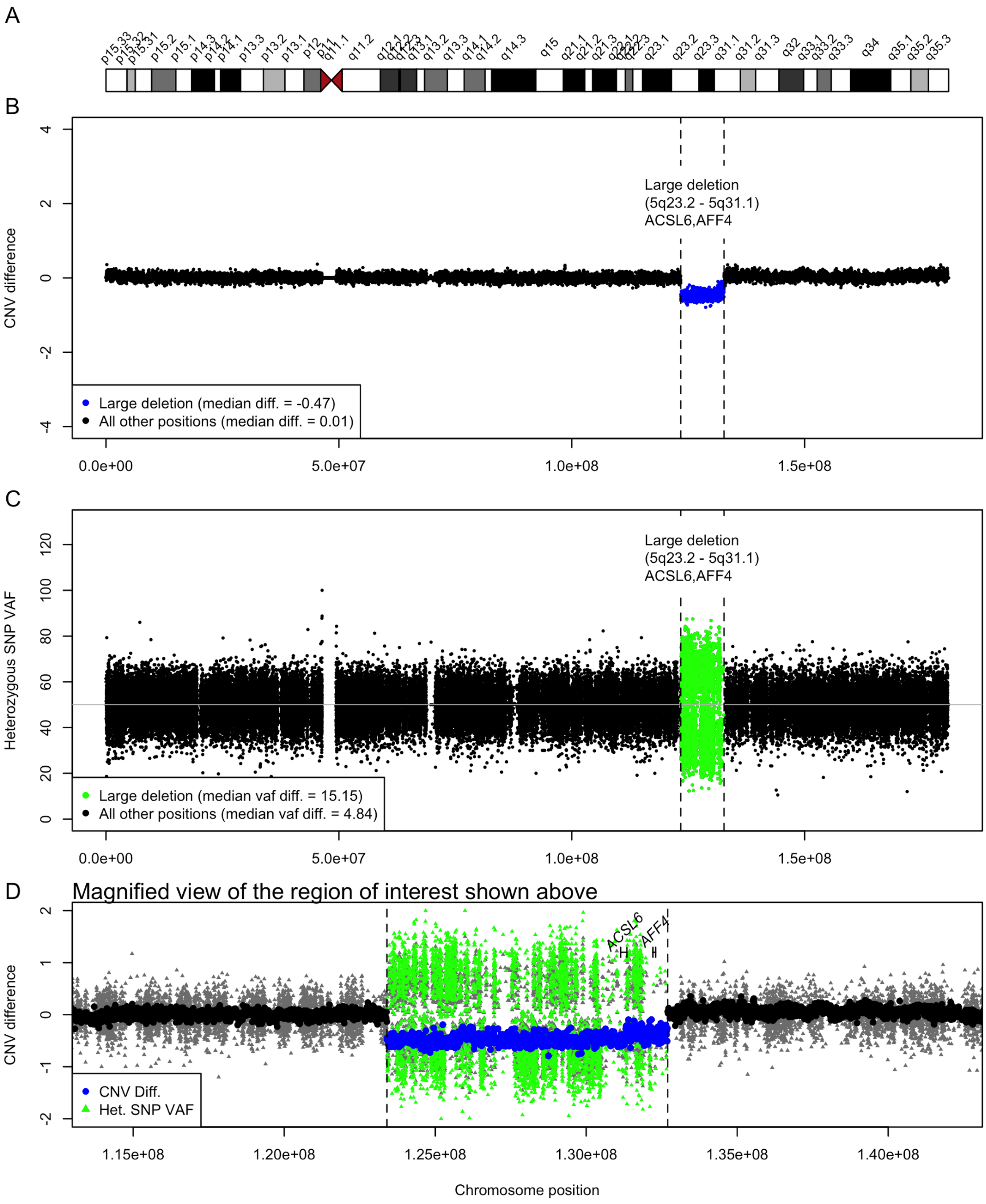


Figure S19. Large deletion at 6q12 - 6q23.2 (affecting six cancer genes)
(A) Ideogram showing Giemsa banding and labels for chromosome 3. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

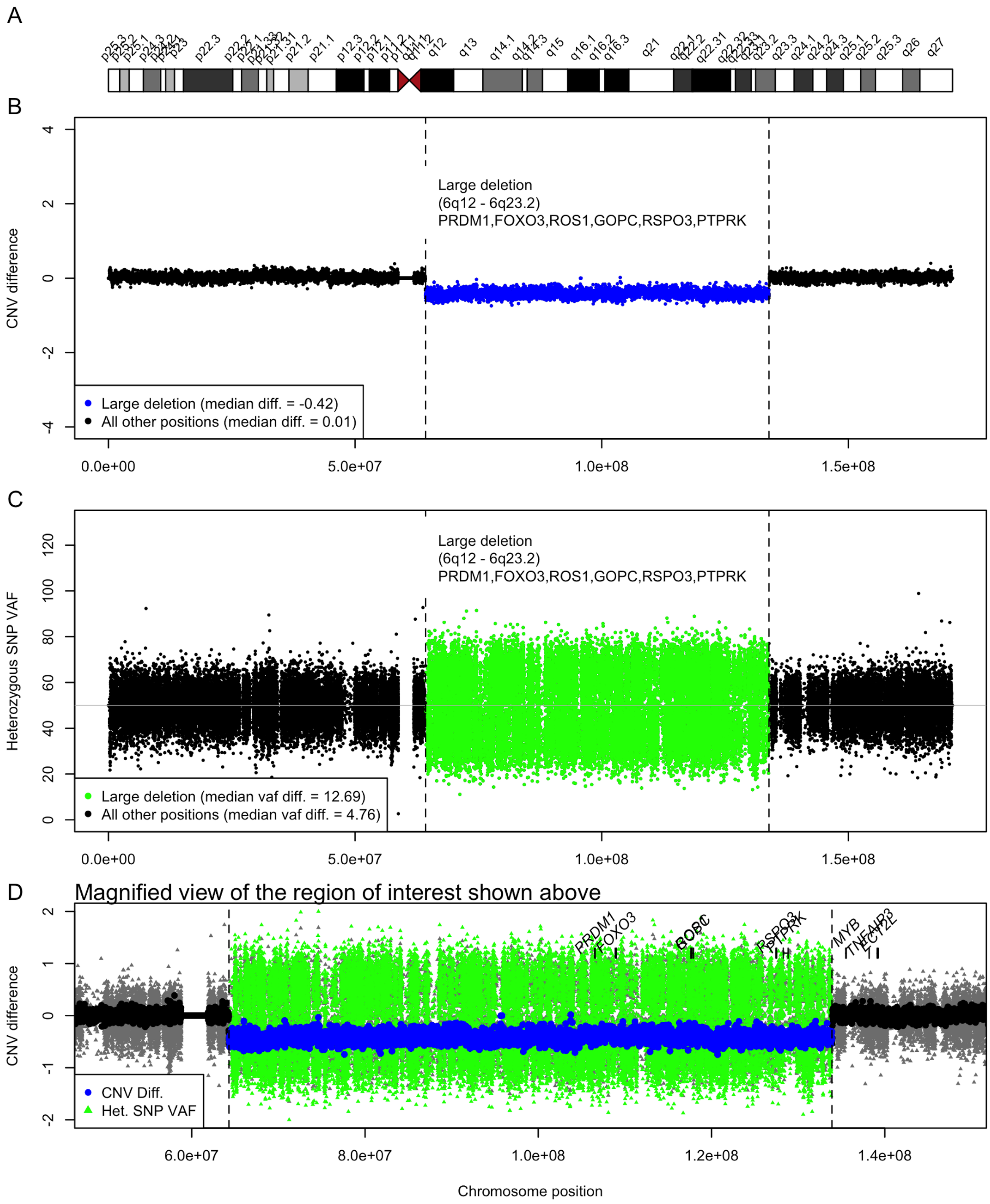


Figure S20. Focal deletion at 7p12.2 - 7p12.1 (affecting *IKZF1* aka Ikaros).
(A) Ideogram showing Giemsa banding and labels for chromosome 7. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

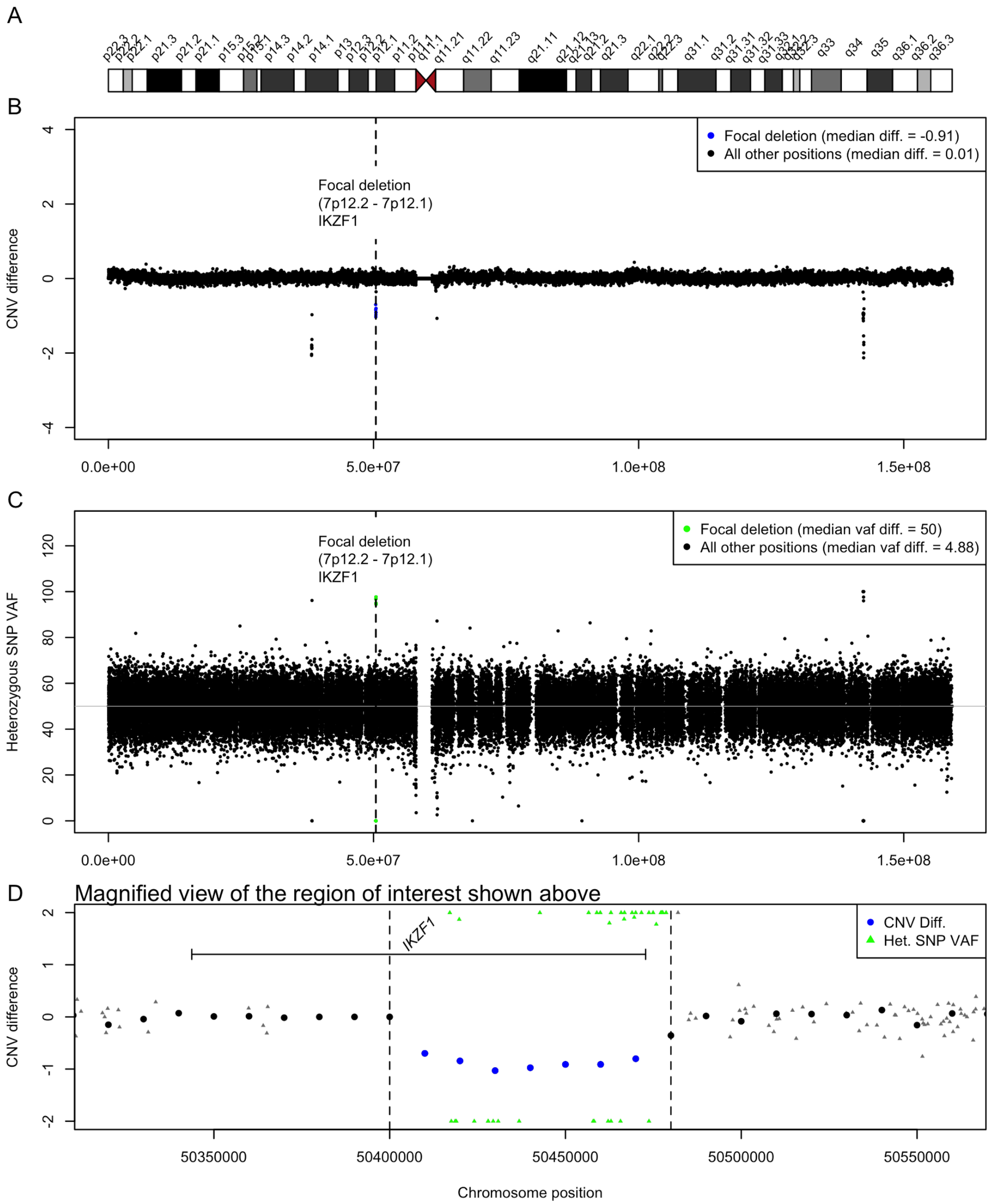


Figure S21. Large deletion at 9p24.3 - 9p13.1 (affecting nine cancer genes including *PAX5*)
(A) Ideogram showing Giemsa banding and labels for chromosome 9. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

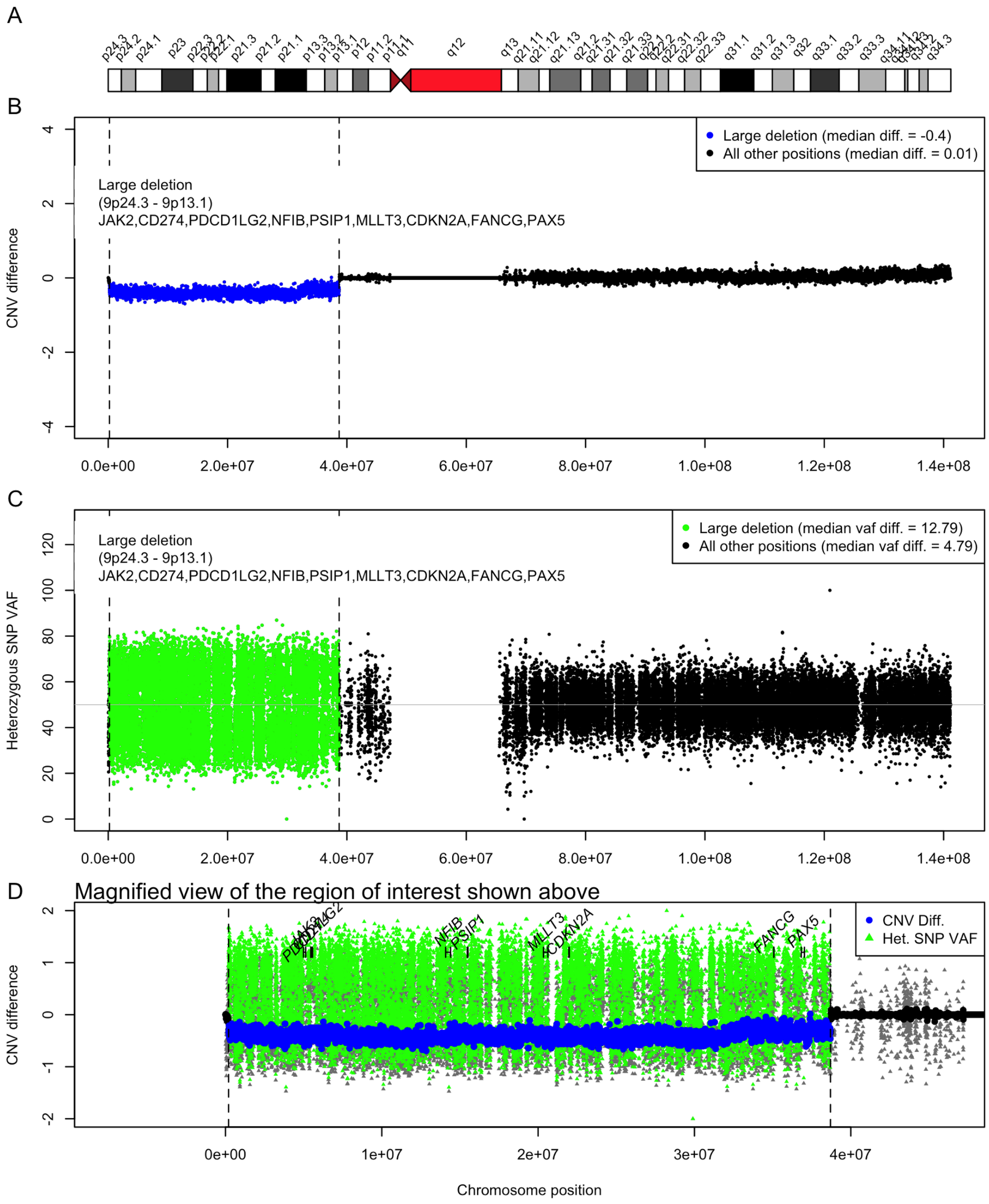


Figure S22. Large deletion at 10q24.1 - 10q24.33 (affecting *TLX1*, *NFKB2*, *SUFU*, and *NT5C2*)
(A) Ideogram showing Giemsa banding and labels for chromosome 10. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

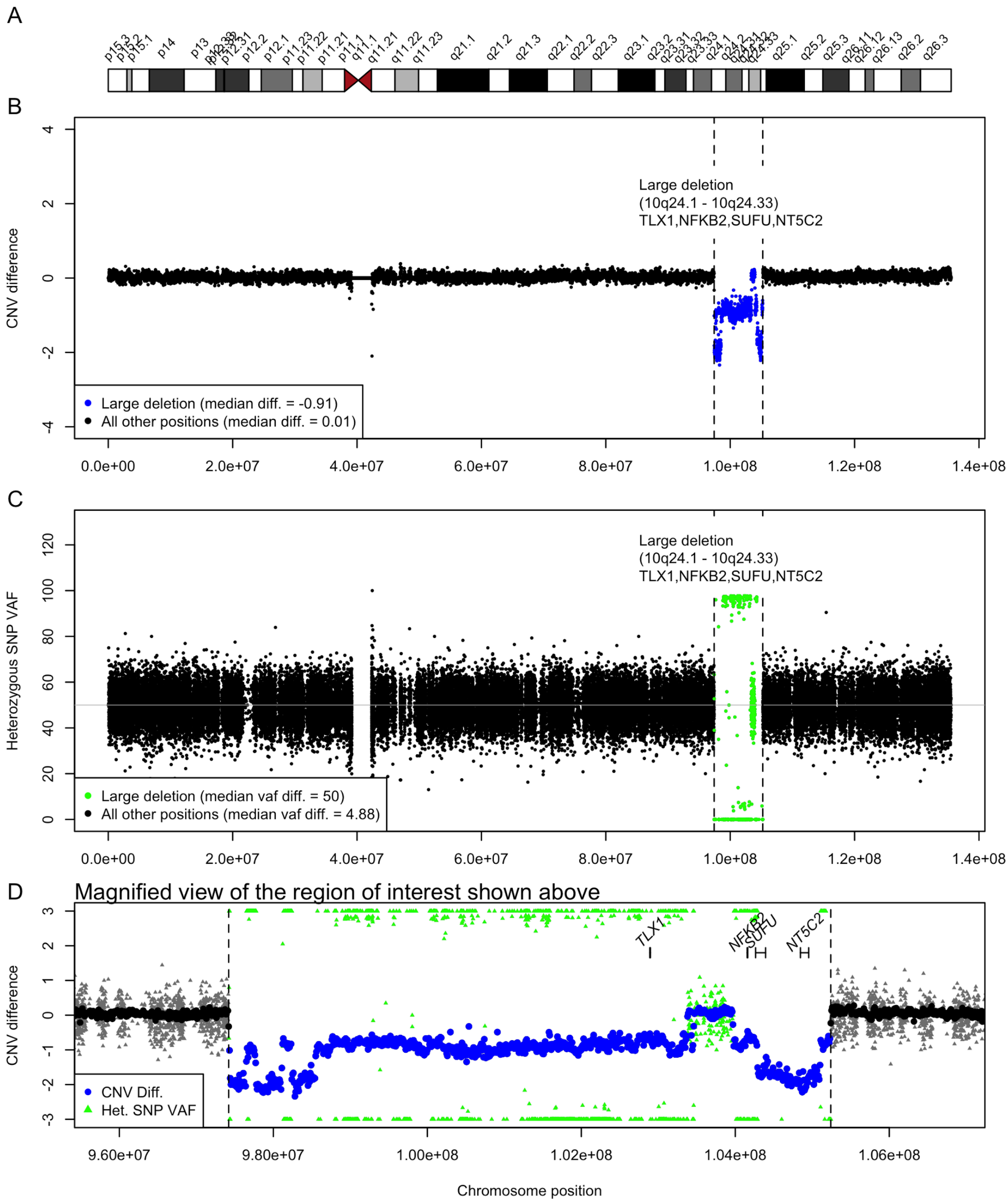


Figure S23. Focal deletion at 11p11.2 (no cancer genes affected)
(A) Ideogram showing Giemsa banding and labels for chromosome 11. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

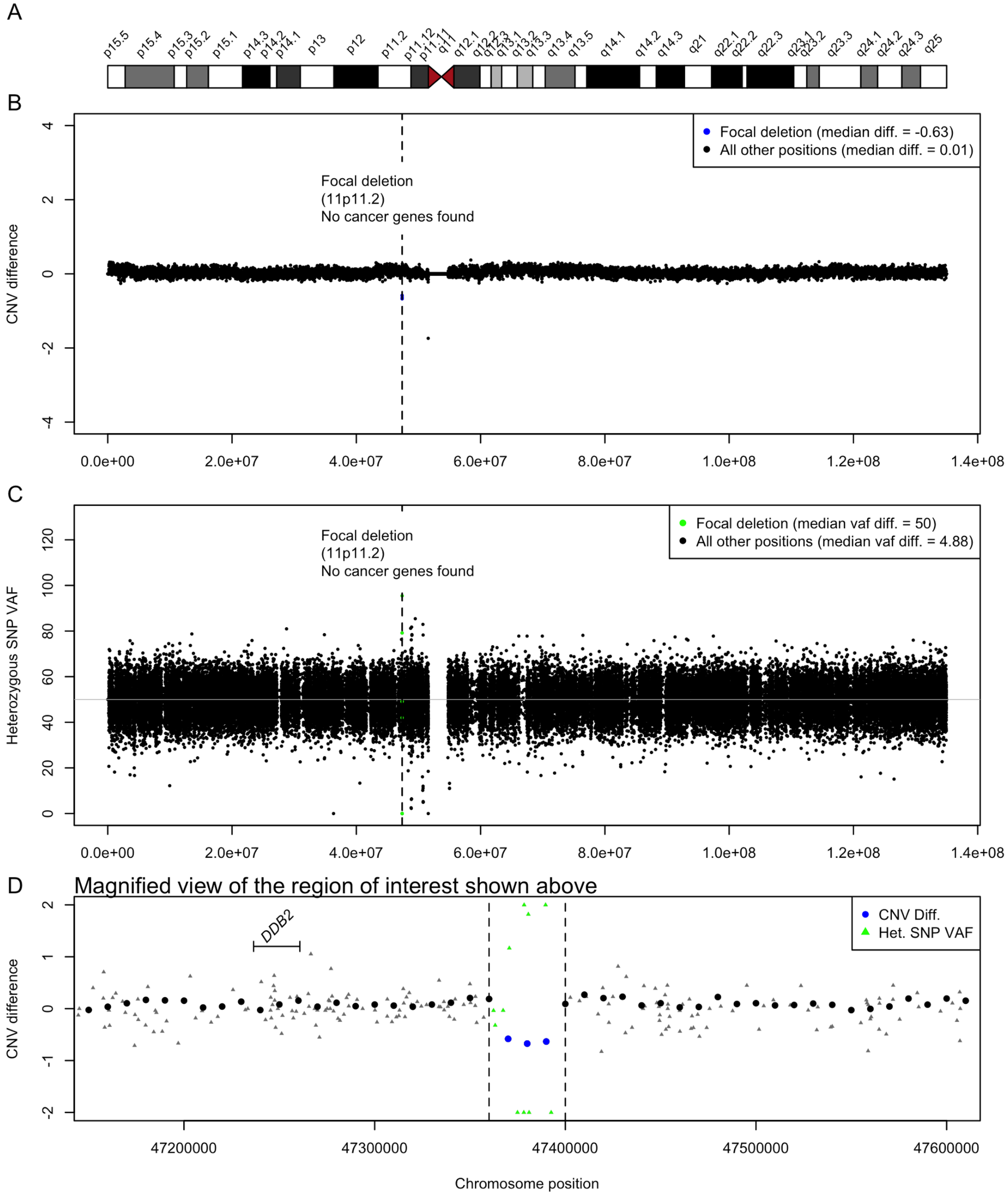


Figure S24. Large deletion at 12p13.2 - 12p12.3 (affecting *ETV6* aka *TEL*, and *CDKN1B*)
(A) Ideogram showing Giemsa banding and labels for chromosome 12. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

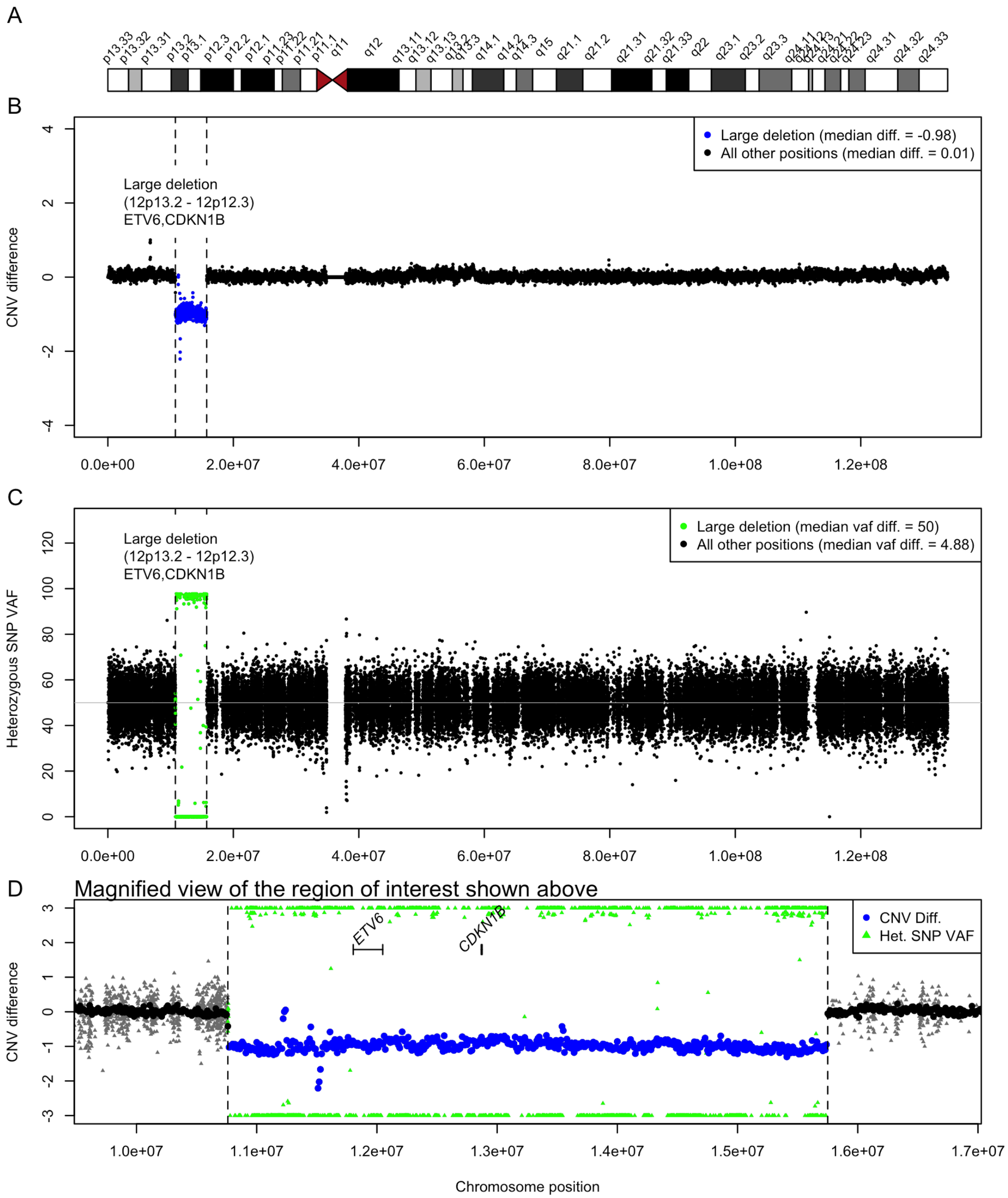


Figure S25. Focal amplification at 12p13.31 (affecting *ZNF384*)
(A) Ideogram showing Giemsa banding and labels for chromosome 12. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

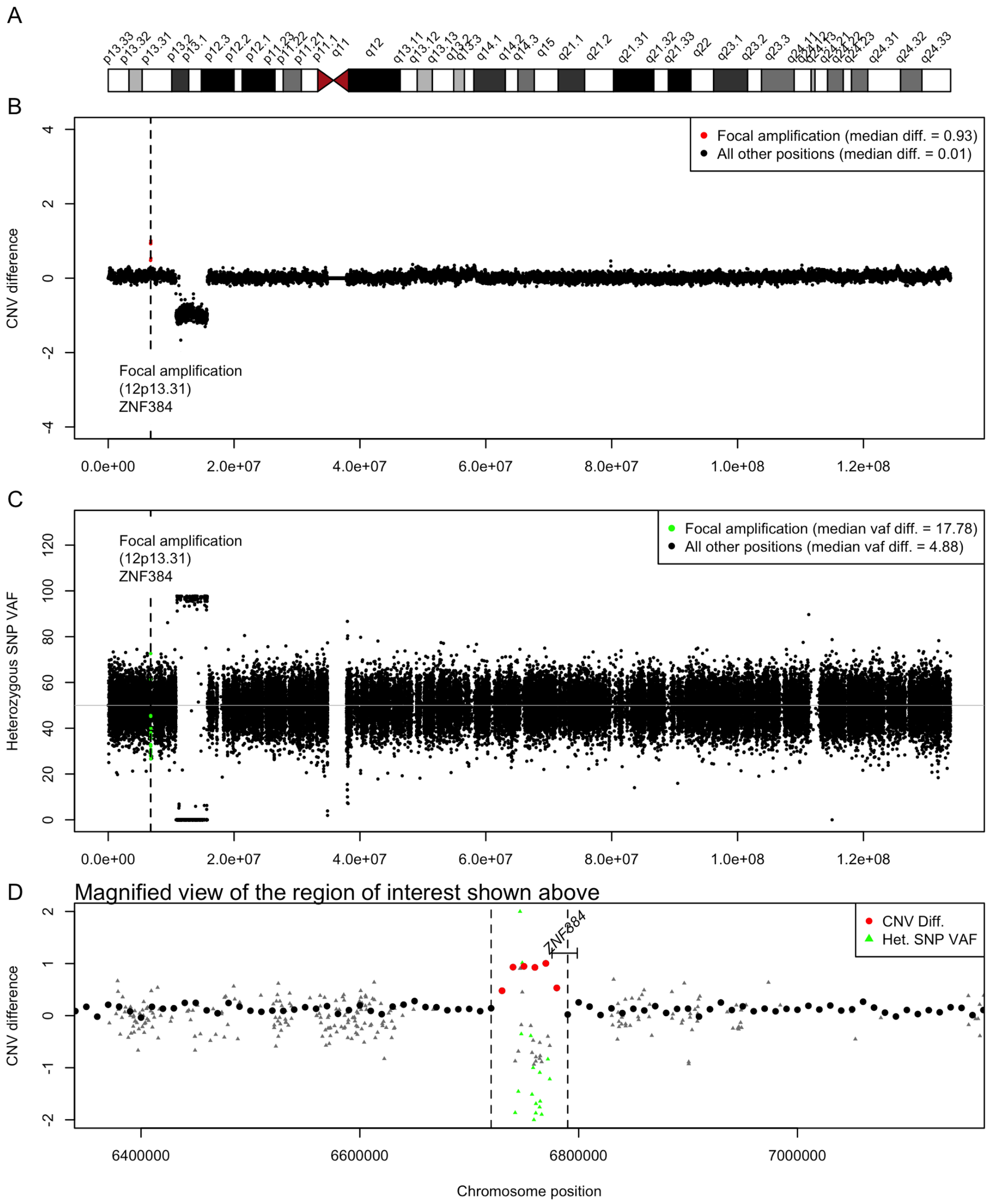


Figure S26. Large deletion at 13q14.13 - 13q14.3 (affecting RB1)
(A) Ideogram showing Giemsa banding and labels for chromosome 13. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

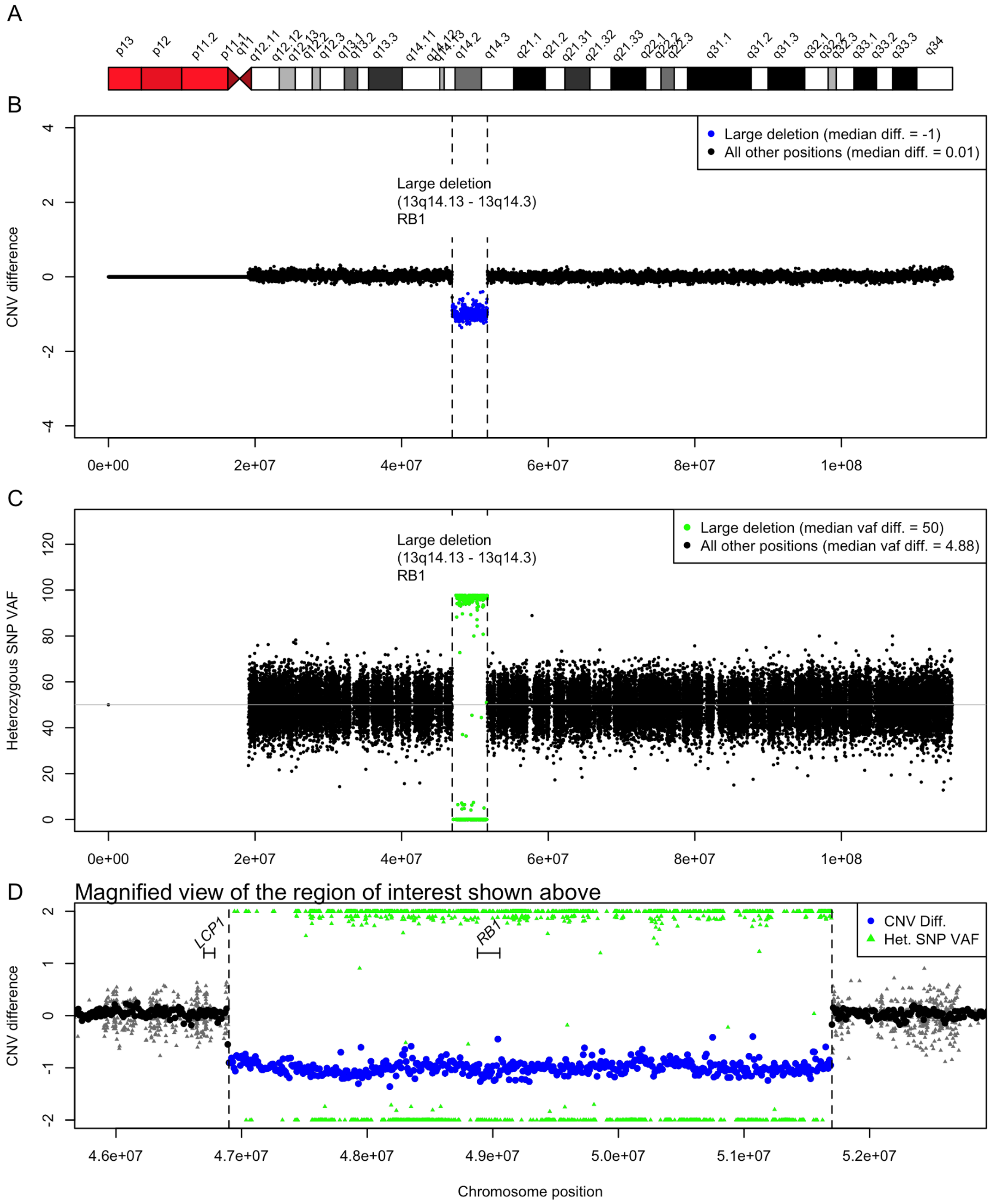


Figure S27. Large region of copy neutral LOH at 17q (30 cancer genes affected)
(A) Ideogram showing Giemsa banding and labels for chromosome 18. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

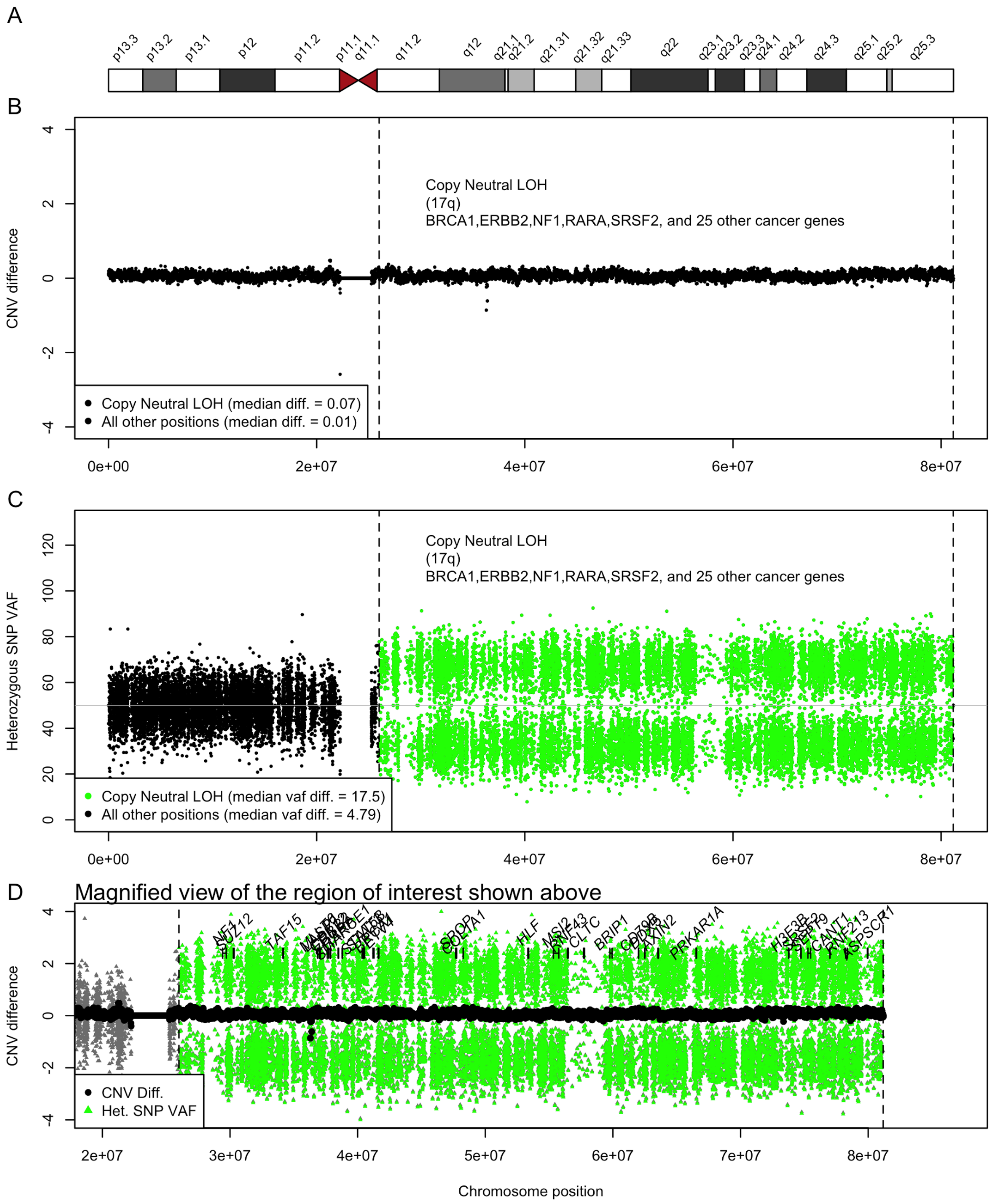


Figure S28. Large deletion at 18p11.32 - 18p11.1 (no cancer genes affected)
(A) Ideogram showing Giemsa banding and labels for chromosome 18. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

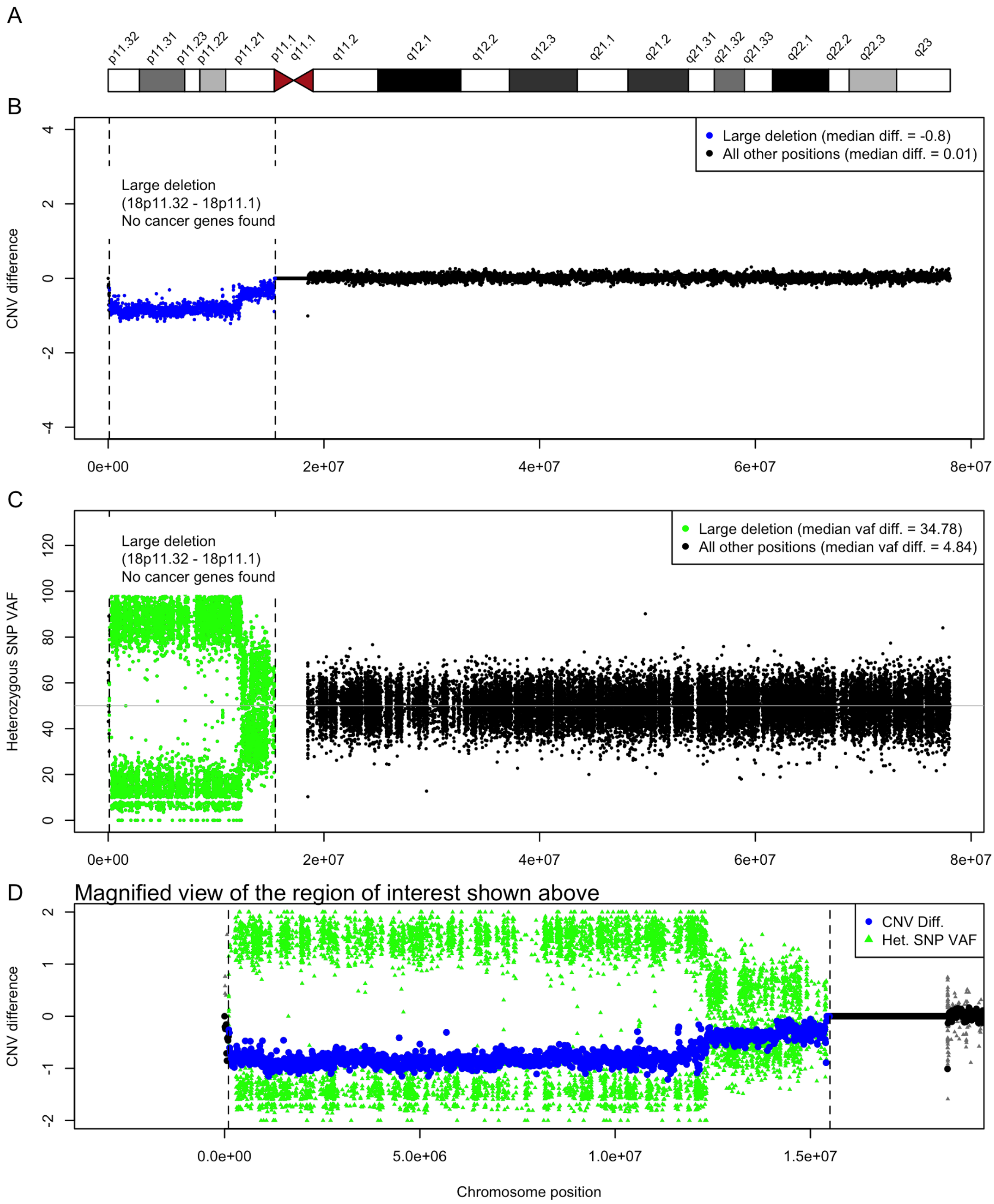


Figure S29. Large deletion at 20p13 - 20p11.1 (No cancer genes affected)
(A) Ideogram showing Giemsa banding and labels for chromosome 20. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

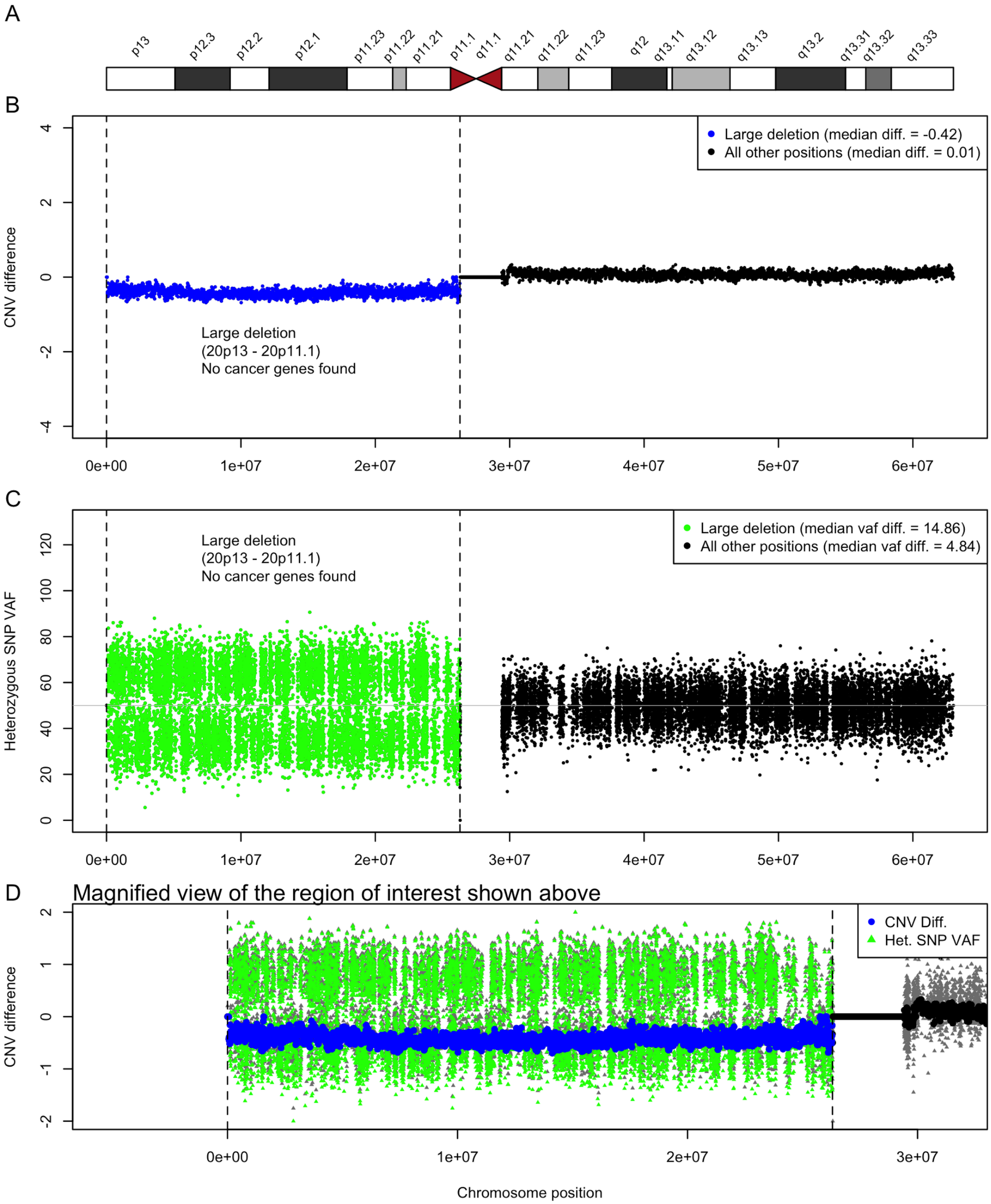


Figure S30. Focal deletion at 22q12.1 (near *CHEK2*)
(A) Ideogram showing Giemsa banding and labels for chromosome 22. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

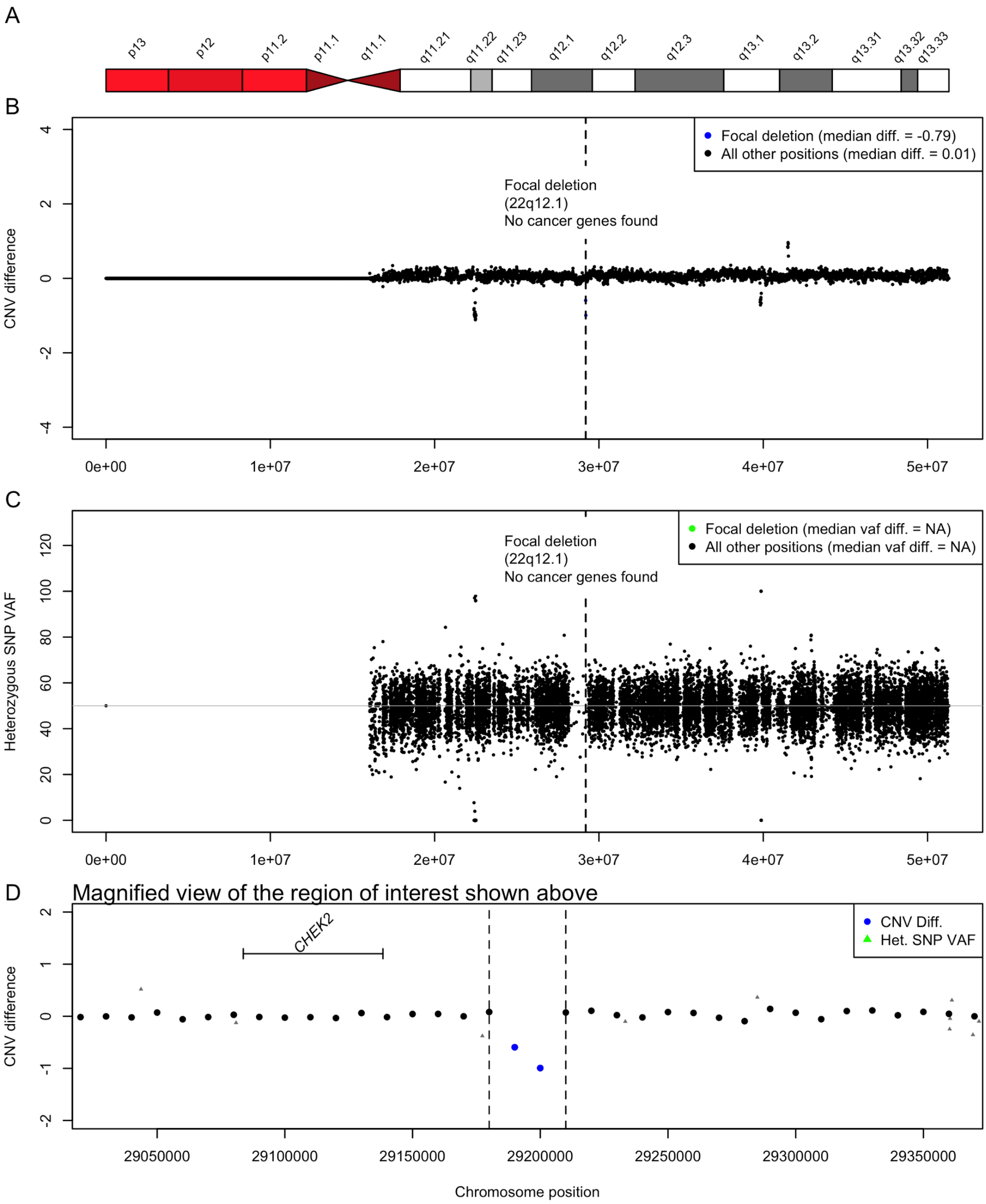


Figure S31. Focal amplification at 22q13.2 (affecting *EP300*)
(A) Ideogram showing Giemsa banding and labels for chromosome 22. (B) CNV differences are displayed for a verified CNV event (see **Table S3** for details) indicated by vertical dotted lines. (C) VAFs for heterozygous SNPs of the patient provide a readout of LOH. (D) CNV differences with LOH VAF data overlaid are plotted for a magnified view of the ROI. Refer to **Figure S17** for a more detailed description of this view.

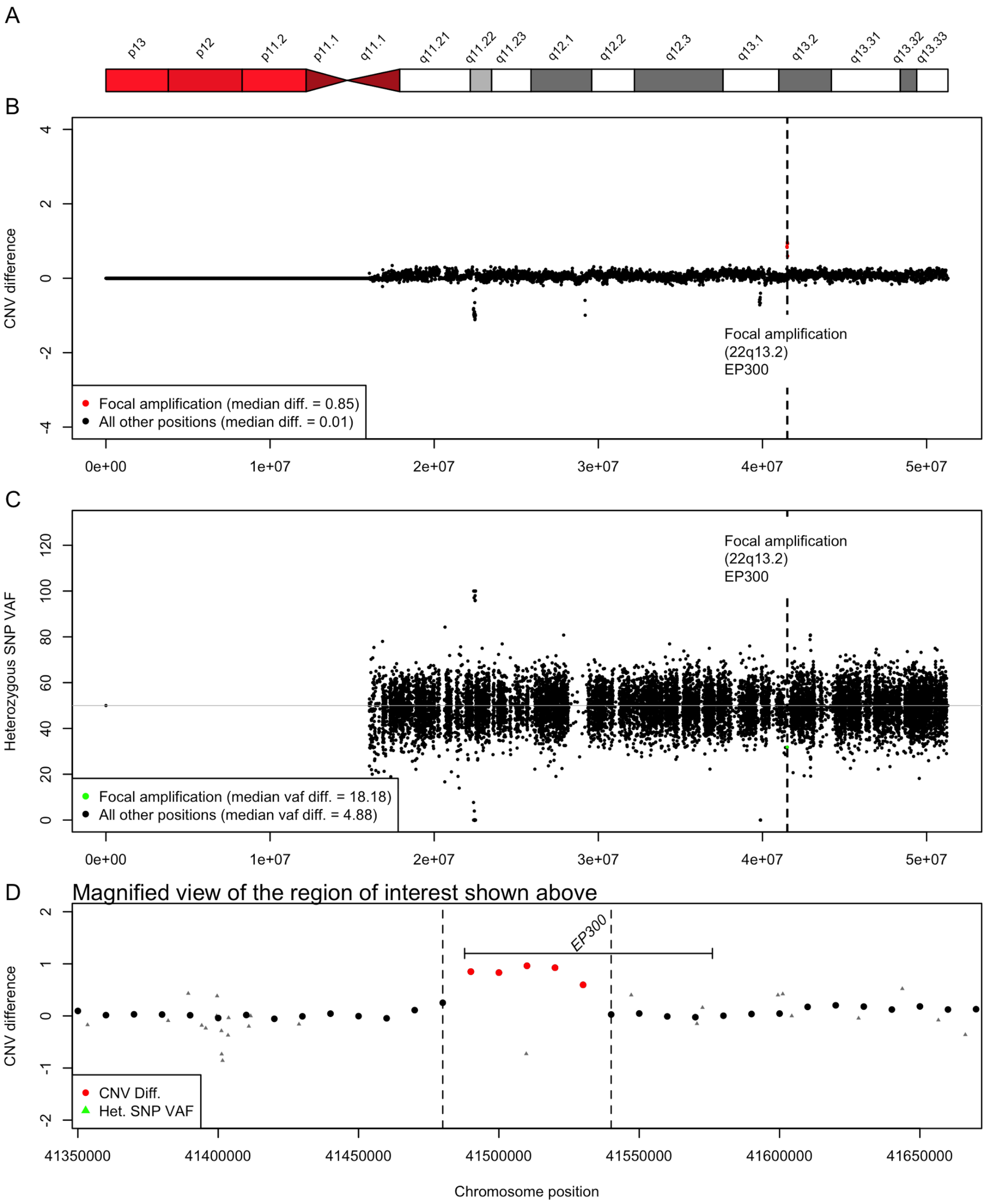


Figure S32. Medium sized deletion affecting *XBP1* exon 1 (12,799 bp deletion, 22:29195159-29207958) Manual review of data supporting a deletion that knocks out the first exon of *XBP1* in the second relapse, sorted blasts tumor sample. (A) An IGV screenshot showing coverage around the predicted deletion breakpoints from the SV caller Manta. Coverage and alignments are shown for the normal skin WGS data followed by second relapse WGS data. (B) An 'svviz' view of the reads aligning to the predicted deletion allele sequence showing WGS reads from the tumor aligning but no reads aligning in the normal WGS data. (C) These alignments are contrasted with reads from both tumor and normal aligning to the reference allele. Since we have relapse 2 reads that align to both the alternate and reference alleles we know that this deletion is heterozygous in the tumor. Since there are no normal reads that align to the alternate (deletion) allele, we know that this deletion is somatic.

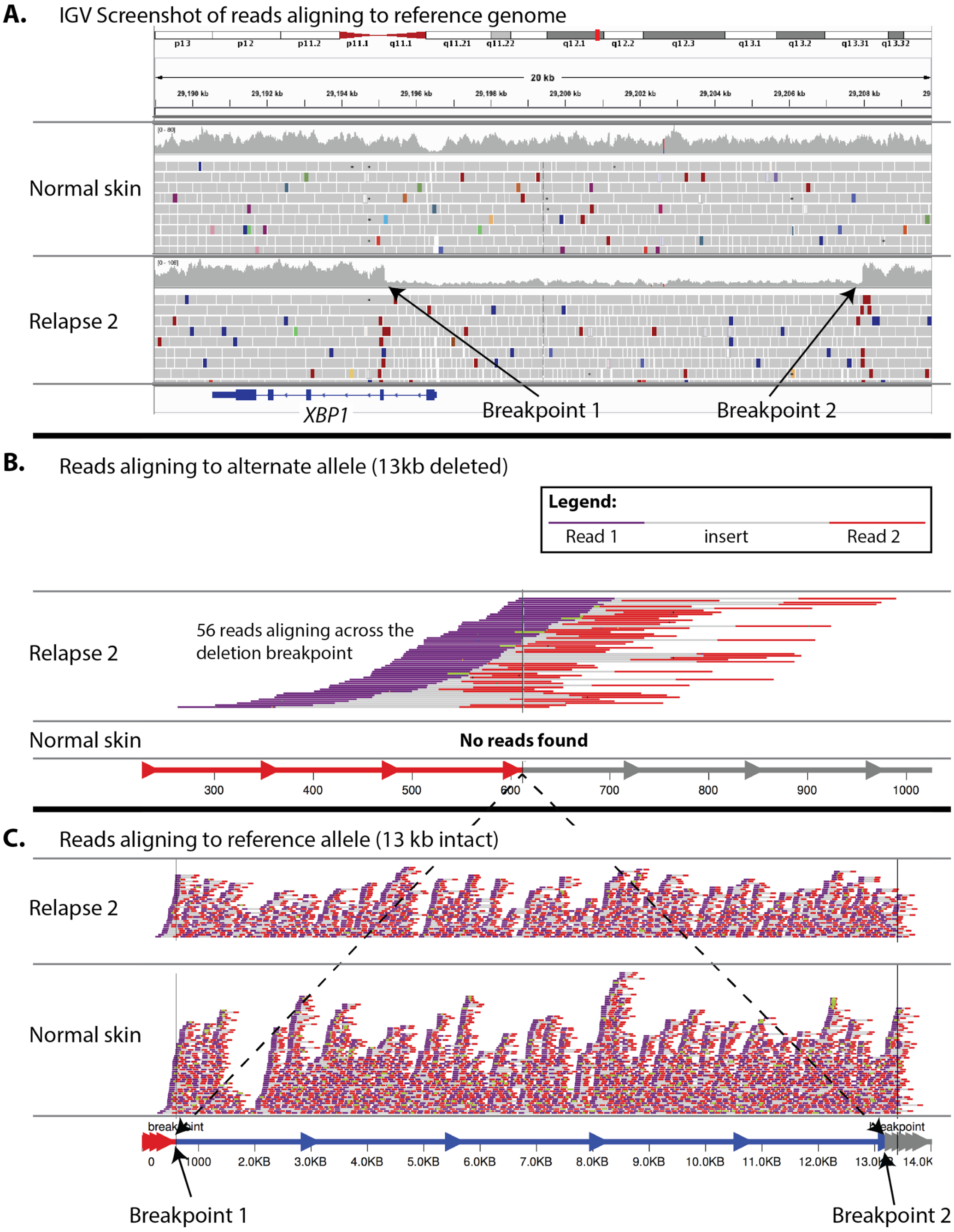


Figure S33. WGS support for discovery of an *EP300-ZNF384* fusion
Supporting WGS reads for the *EP300-ZNF384* translocation. (A) Discordant reads that encompass the translocation identified using pairoscope (http://pairoscope.sourceforge.net/). No encompassing reads were identified in the normal sample indicating a somatic event. (B) Reads spanning the translocation allele identified by svviz. No reads supporting the translocation were found in the normal sample indicating a somatic event. (C) Reads aligning to the reference allele at the chr12, chr22 breakpoints identified by svviz. Presence of reads indicate that the tumor sample is heterozygous for the translocation allele at both breakpoints. The normal sample is homozygous for the reference allele at both breakpoints.

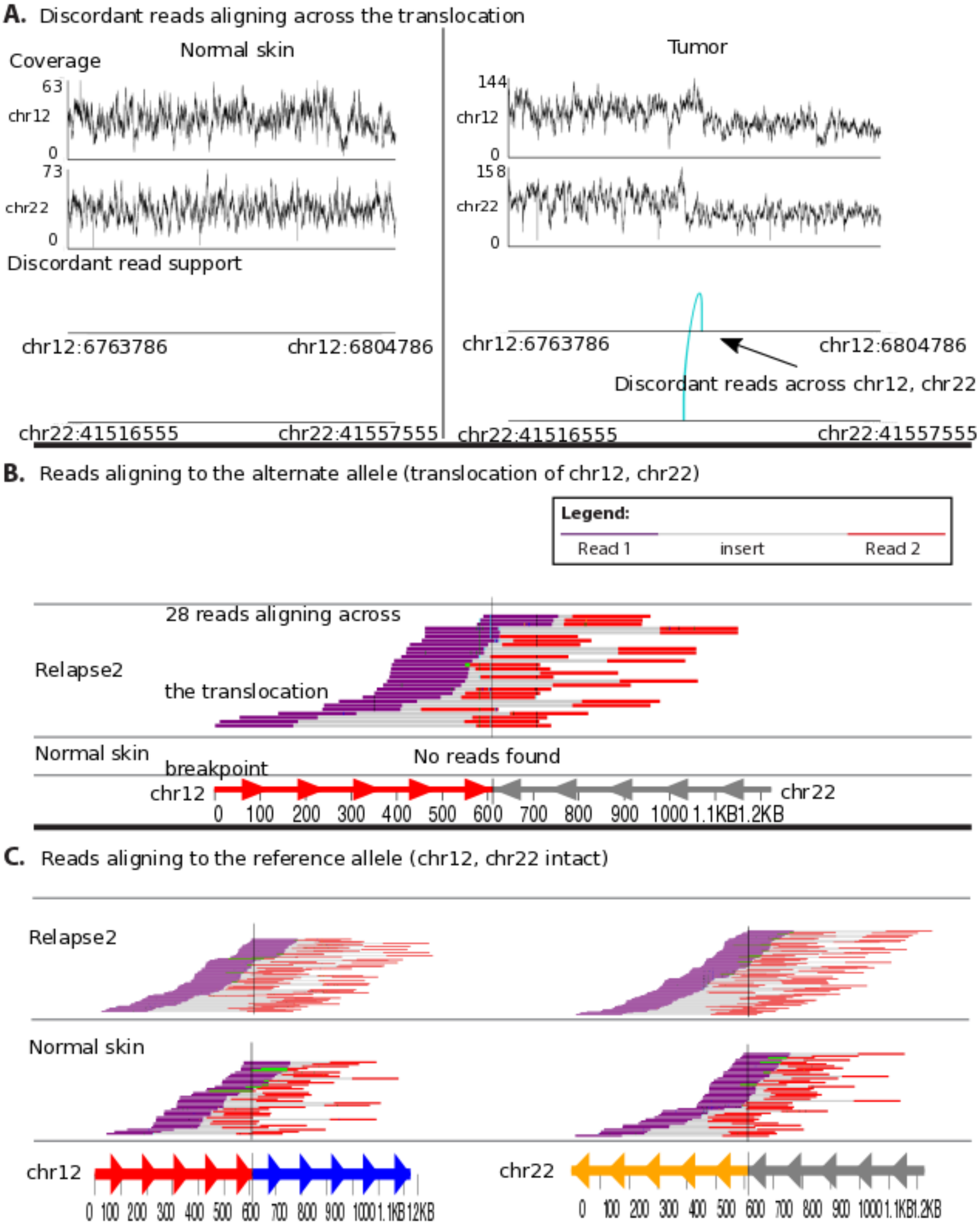
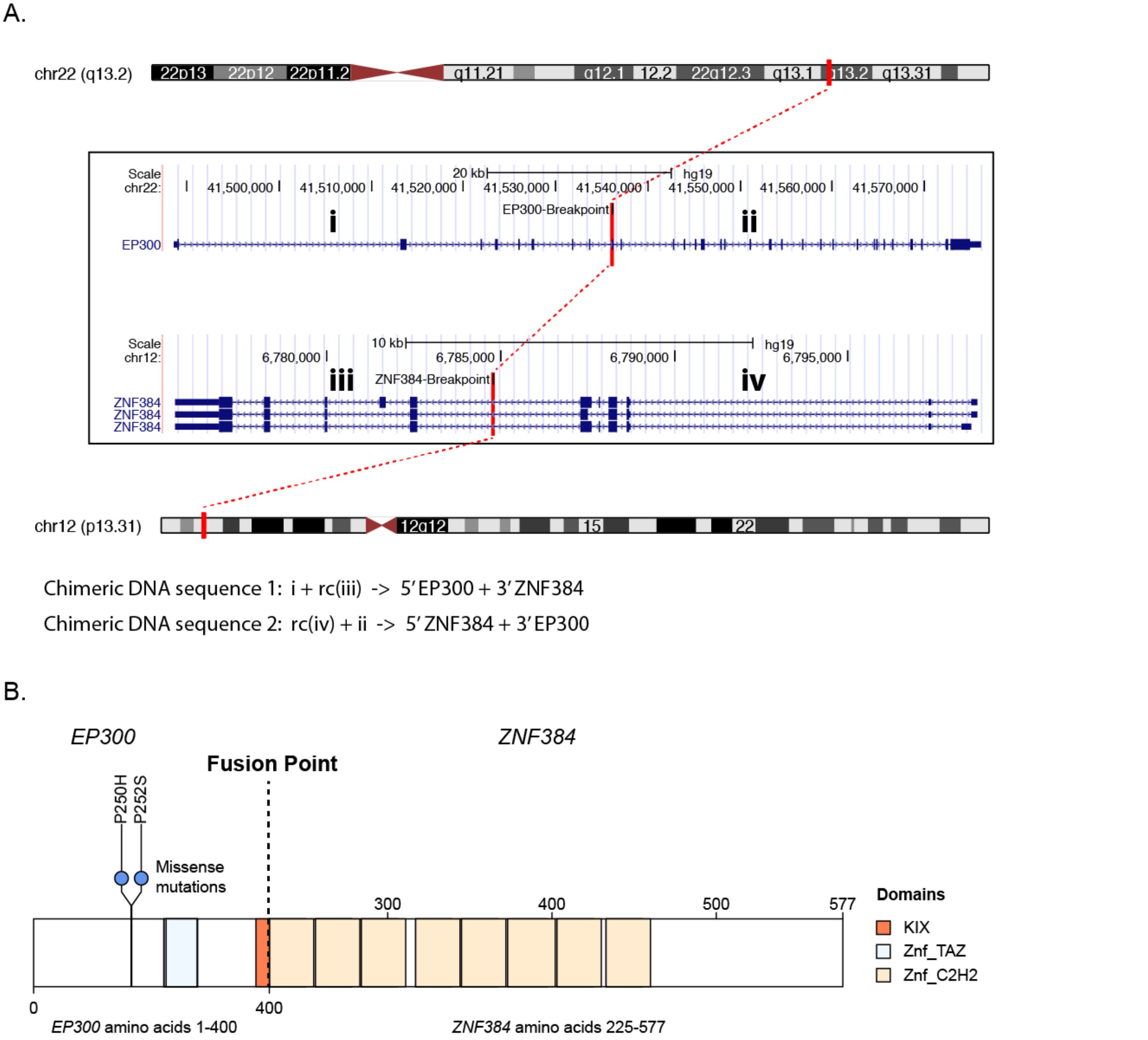


Figure S34. *EP300-ZNF384* fusion, predicted DNA and protein structures (with 2 missense mutations) (A) The DNA breakpoints for a predicted *EP300-ZNF384* fusion [t(12;22)(p13;q13)] are depicted in relation to the position of exons for known transcripts (blue lines and rectangles). The breakpoint position is linked back to ideograms for each chromosome by red dotted lines. *EP300* is transcribed from the positive strand (left to right), while *ZNF384* is transcribed from the negative strand (right to left). Both breakpoints occur within introns of *EP300* and *ZNF384,* respectively. The predicted arrangement of fused chromosomes assuming a reciprocal event are depicted below as 'Chimeric DNA sequence 1' and 'Chimeric DNA sequence 2'. (B) The fusion protein sequence predicted by full length cDNA cloning of the EP300-ZNF384 transcript is depicted with known protein domains indicated as colored rectangles. The predicted fusion is 577 amino acids in length. The position of two missense mutations detected within the EP300 gene (in linkage with the translocation event) are depicted as blue circles.

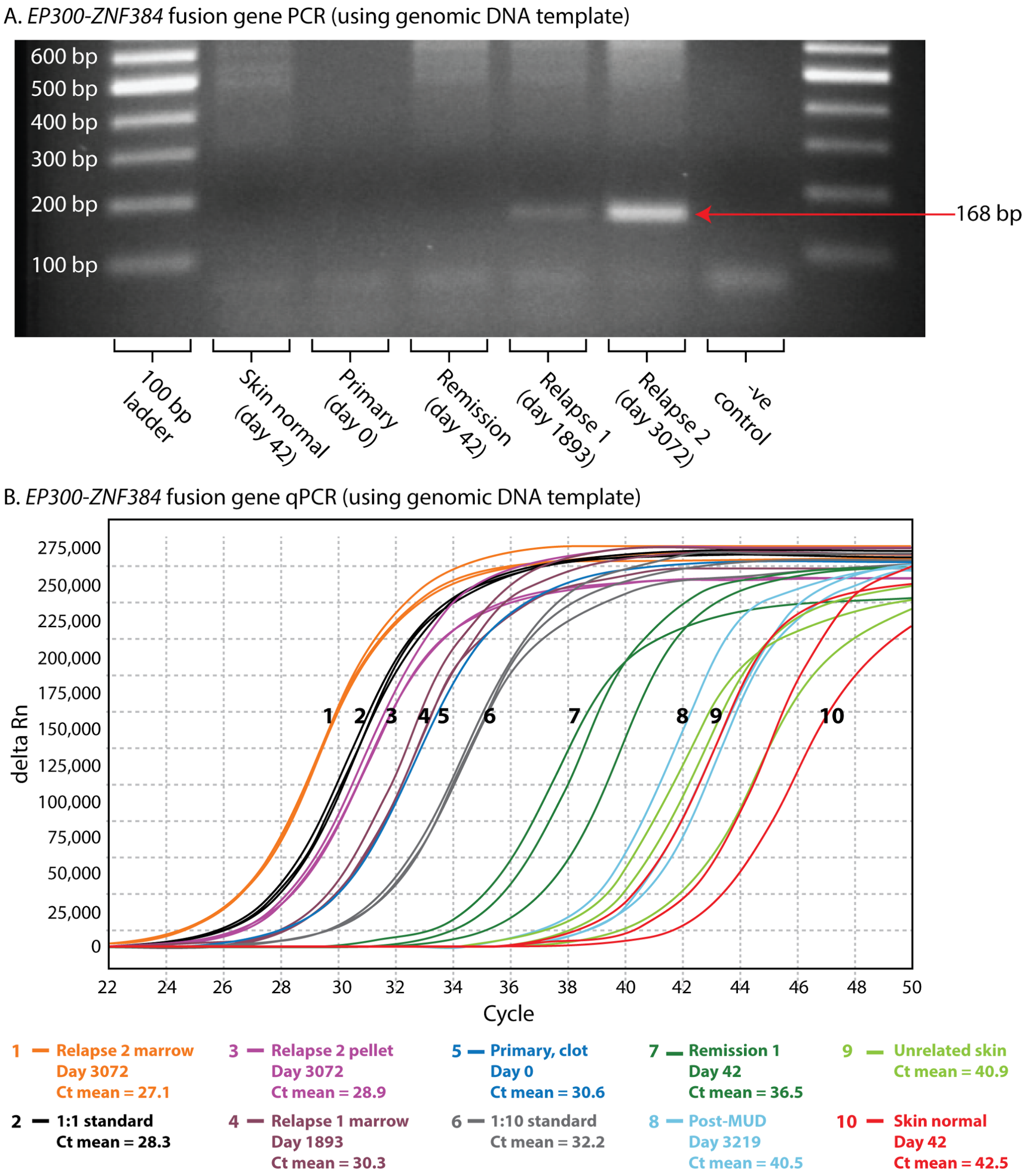


Chimeric DNA sequence 1: i + rc(iii) -> 5' EP300 + 3' ZNF384

Chimeric DNA sequence 2: rc(iv) + ii -> 5' ZNF384 + 3' EP300

Figure S35. PCR and qPCR support for the ALL1 *EP300-ZNF384* fusion

Two assays of genomic DNA for an *EP300-ZNF384* translocation. (A) An RT-PCR assay was used to amplify a 168 bp amplicon representing the translocation breakpoint. Primers were selected in the introns of *EP300* and *ZNF384*, adjacent to the genomic DNA breakpoint (see **Methods** for primer sequences and assay details). This assay was applied to bone marrow samples obtained from our patient at four time points (B) A qPCR assay was conducted with the same primer sequences as the RT-PCR assay. This assay was applied to seven samples from our patient along with positive and negative controls. Colored lines indicate measured fluorescence from Sybr green incorporation into PCR product plotted against PCR cycle number (see **Methods** for additional assay details).



A. *EP300-ZNF384* fusion gene PCR (using genomic DNA template)

B. *EP300-ZNF384* fusion gene qPCR (using genomic DNA template)

Figure S36. ***TBX19-SUFU*** **DNA translocation**

Supporting WGS reads for the *TBX19-SUFU* translocation. (A) Discordant reads that encompass the translocation identified using pairoscope. No encompassing reads were identified in the normal sample indicating a somatic event. (B) Reads spanning the translocation allele identified by svviz. No reads supporting the translocation were found in the normal sample indicating a somatic event. (C) Reads aligning to the reference allele at the chr1 and chr10 breakpoints identified by svviz. No reads were identified in the tumor at the chr10 breakpoint indicating loss of the reference allele in this region. The chr1 breakpoint is heterozygous in the tumor. The normal sample is homozygous for the reference allele at both breakpoints.

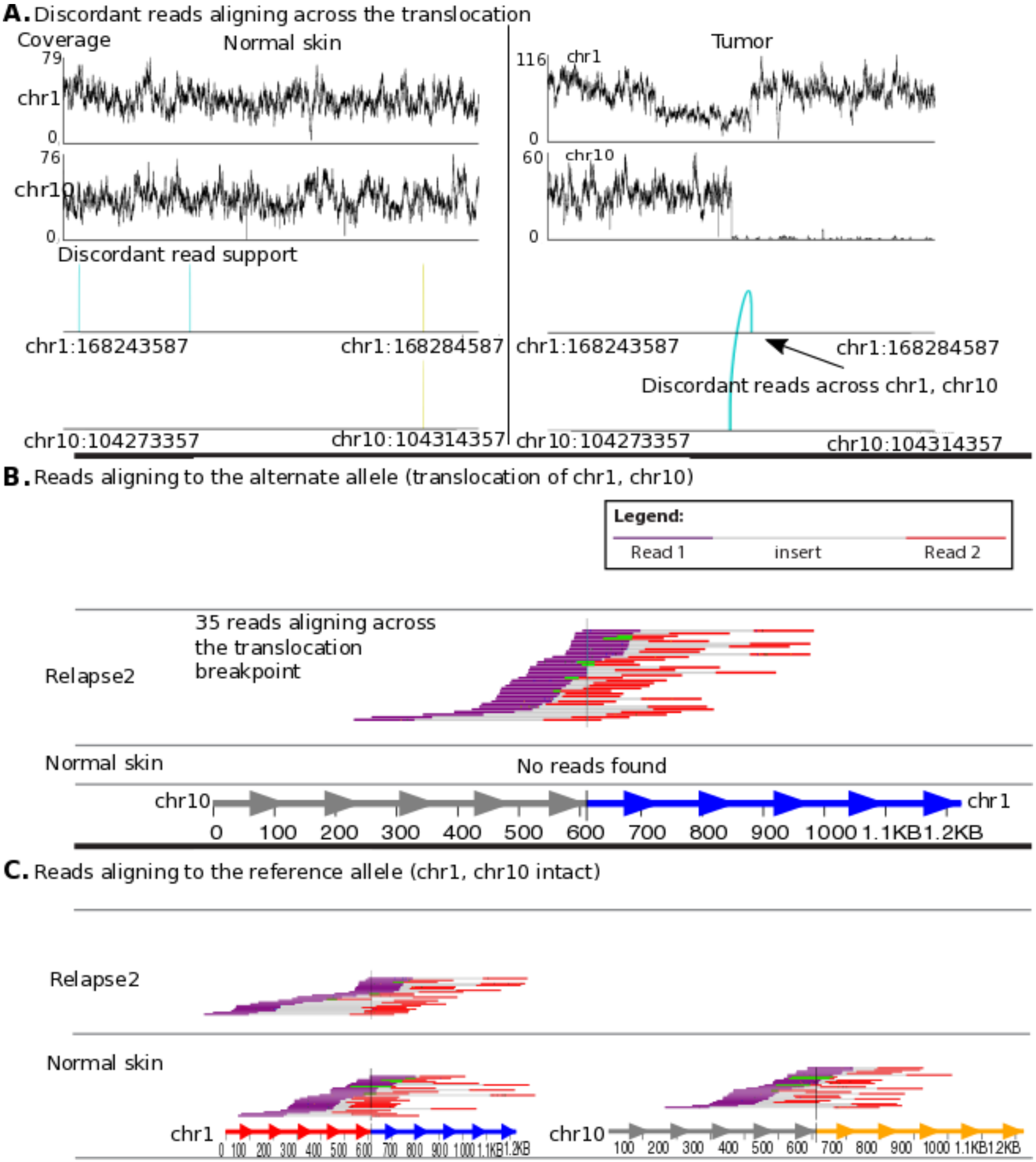


Figure S37. ***ADCY10-CC2D2B* DNA translocation**

Supporting WGS reads for the *ADCY10-CC2D2B* translocation. (A) Discordant reads that encompass the translocation identified using pairoscope. No encompassing reads were identified in the normal sample indicating a somatic event. (B) Reads spanning the translocation allele identified by svviz. No reads supporting the translocation were found in the normal sample indicating a somatic event. (C) Reads aligning to the reference allele at the chr1 and chr10 breakpoints identified by svviz. Only one read is identified in the tumor at the chr10 breakpoint indicating loss of the reference allele in this region. The chr1 breakpoint is heterozygous in the tumor. The normal sample is homozygous for the reference allele at both breakpoints.

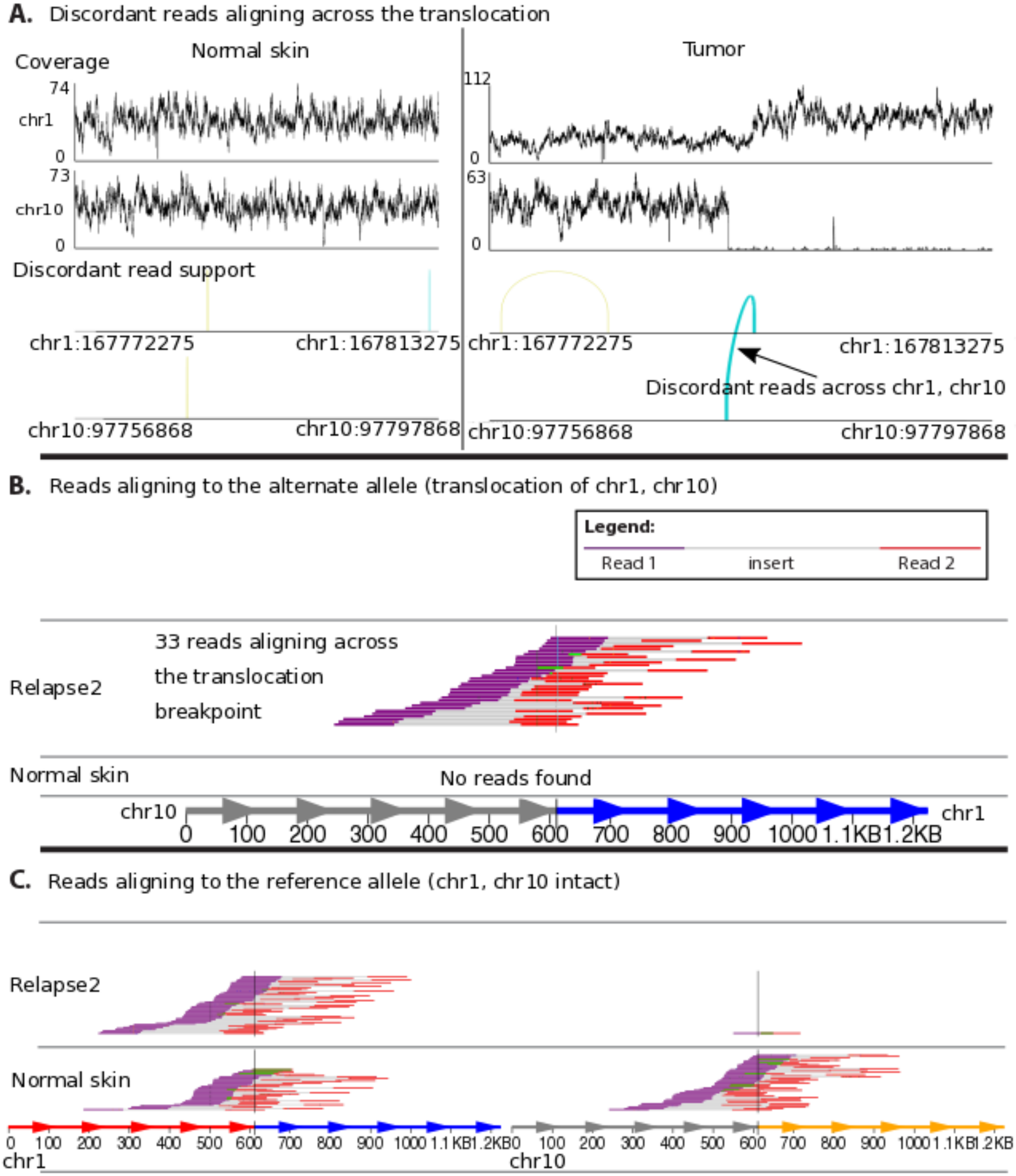


Figure S38. RNA expression of somatic and germline variants of the second relapse
Scatter plots compare the VAF observed in genomic DNA (WGS/Exome/Capture) and RNA samples obtained at second relapse (day 3,072). Each point represents a single variant. Somatic variants are plotted in panel A, while germline are plotted in panel B. An RNA gene expression value is represented on a color scale for each point where red indicates high expression for the gene harboring the variant and yellow indicates low expression. RNA expression was measured as fragments per kilobase of transcript per million fragments (FPKM) with a linear transform applied such that the minimum FPKM was set to 1 (followed by log2 transformation). Only variants predicted to affect the protein coding sequence of a gene are shown. Selected genes are labeled with official HUGO gene symbols.

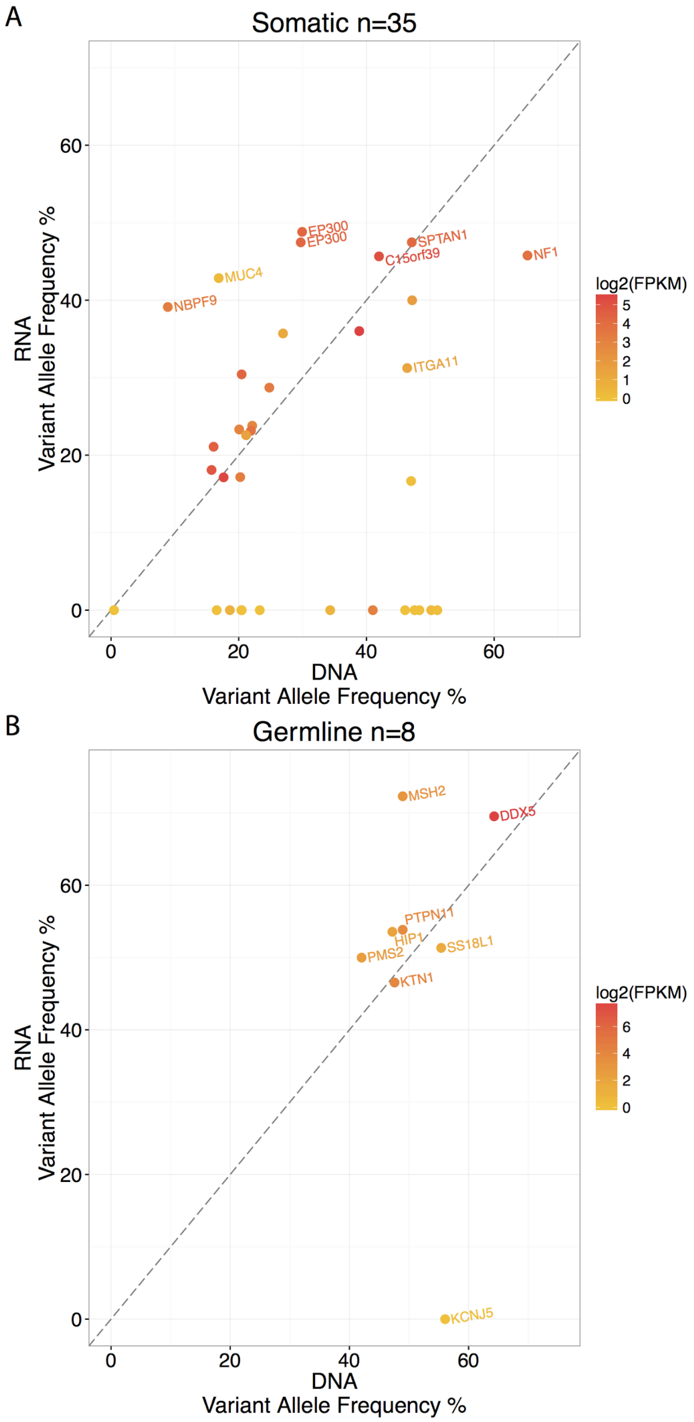
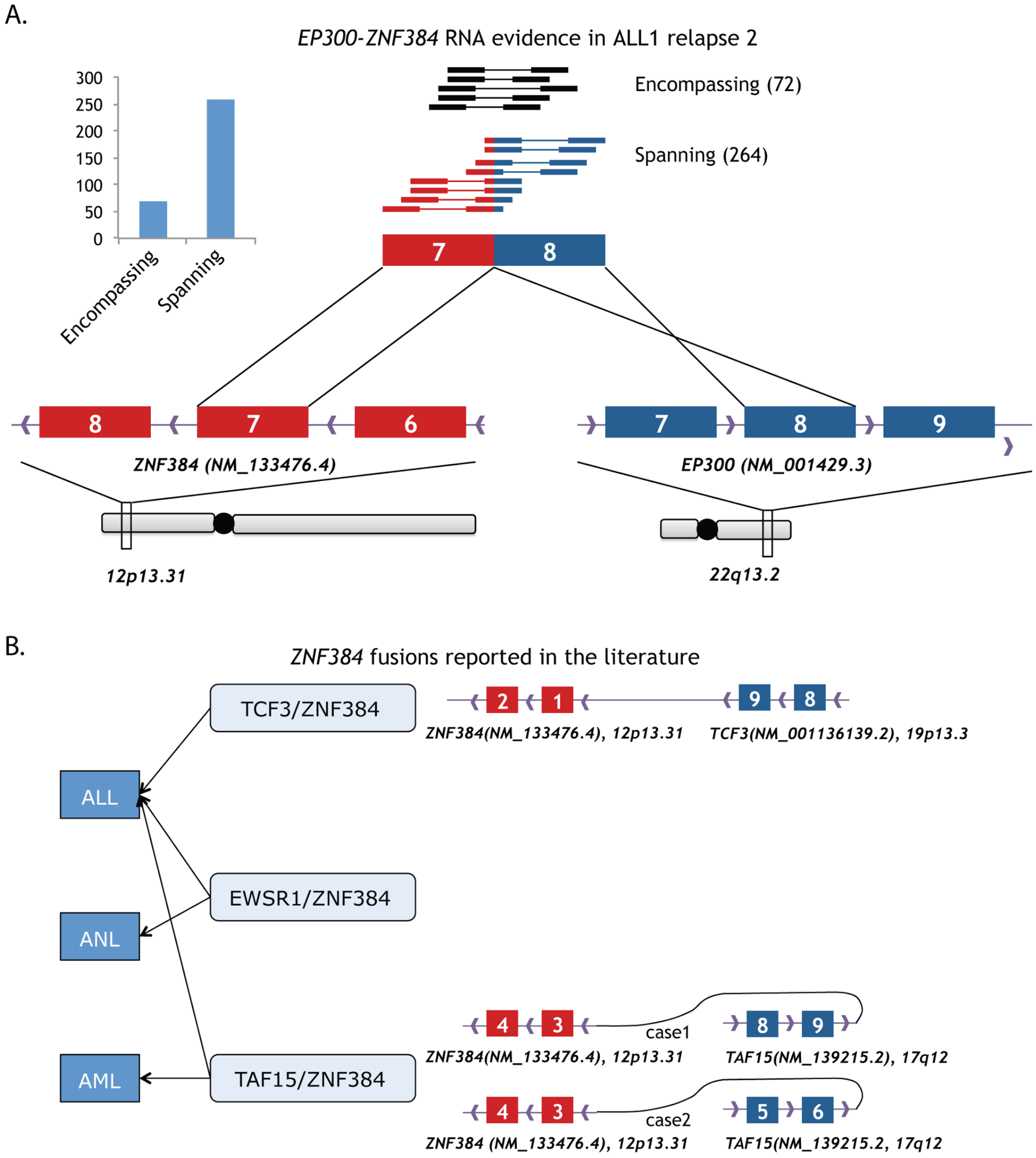


Figure S39. RNA-seq support for the ALL1 *EP300-ZNF384* fusion and literature support for *ZNF384* fusions in leukemia

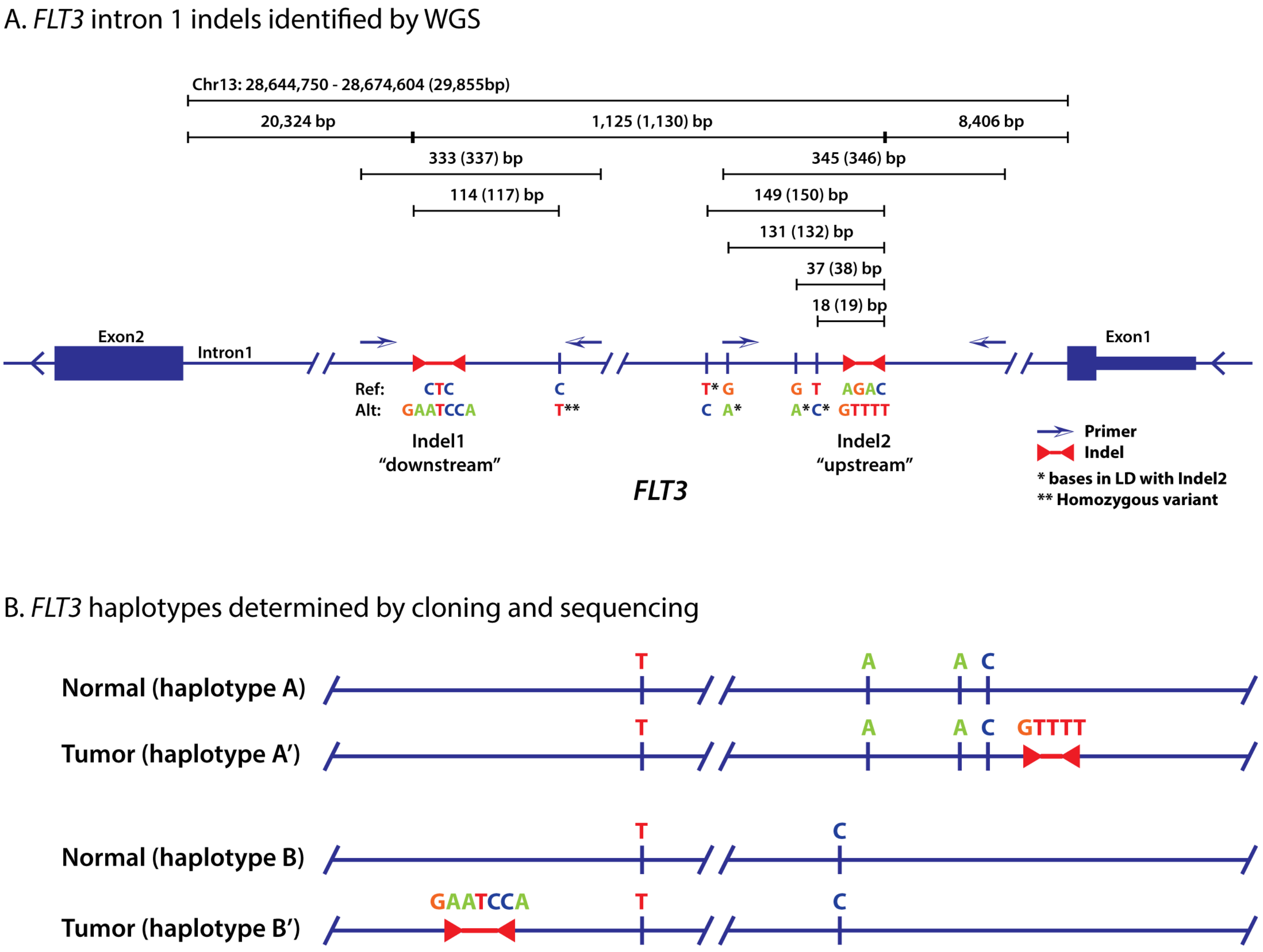
(A) Summary of RNA-seq evidence obtained from the second relapse (day 3,072, sorted blasts) for an *EP300-ZNF384* gene fusion identified by Chimerascan (**Supplementary Methods**). The predicted structure of the fusion based on spanning read support is also shown. (B) Summary of literature support for the occurrence in leukemia of fusions involving *ZNF384* as a 3' gene partner.


A.
EP300-ZNF384 RNA evidence in ALL1 relapse 2
300
250
200
150
100
50
0
Encompassing
Spanning
Encompassing (72)
Spanning (264)
7
8
8
7
6
7
8
9
ZNF384 (NM_133476.4)
EP300 (NM_001429.3)
12p13.31
22q13.2
B.
ZNF384 fusions reported in the literature
TCF3/ZNF384
ALL
EWSR1/ZNF384
ANL
AML
TAF15/ZNF384
2
1
9
8
ZNF384(NM_133476.4), 12p13.31
TCF3(NM_001136139.2), 19p13.3
4
3
case1
8
9
ZNF384(NM_133476.4), 12p13.31
TAF15(NM_139215.2), 17q12
4
3
case2
5
6
ZNF384 (NM_133476.4), 12p13.31
TAF15(NM_139215.2, 17q12

ALL: Acute lymphoblastic leukemia/lymphoblastic lymphoma
ANL: Acute undifferentiated leukemia
AML: Acute myeloblastic leukemia without maturation (FAB type M1)

Figure S40. *FLT3* intron 1 somatic indels

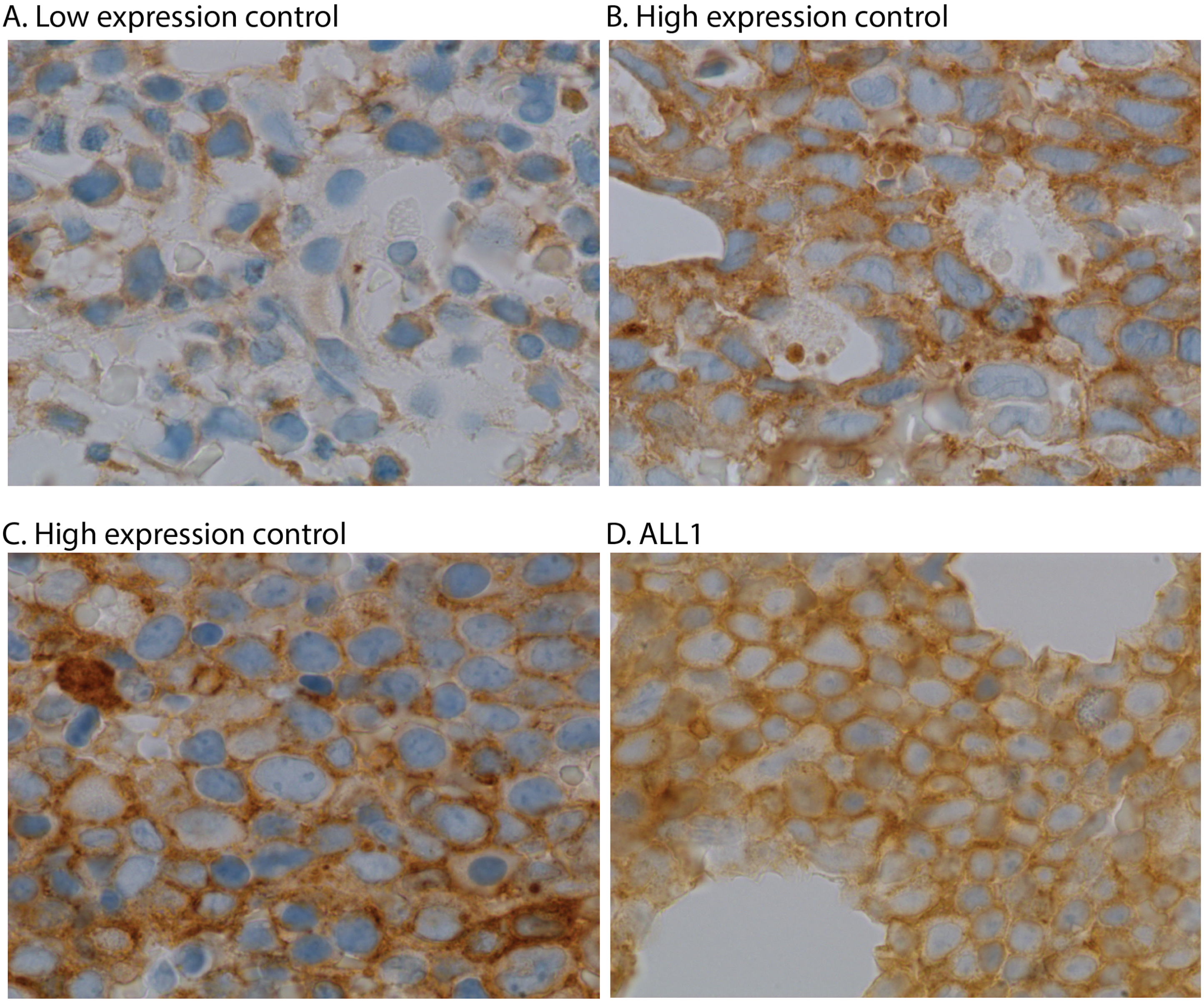
(A) A schematic depiction of two somatic small indels, each a complex substitution is provided. Exon 1 and 2 of *FLT3* are depicted (not to scale) with key features relevant to these somatic events. The somatic mutations themselves are indicated as ‘Indel1’ and ‘Indel2’. The position of primers used to validate these mutations by Sanger sequencing are indicated as blue arrows. The sequences for reference (‘Ref’) and alternative (‘Alt’) alleles are indicated below the *FLT3* gene model. Five SNPs used to establish the haplotypes associated with each somatic mutation are also indicated below the *FLT3* gene model. The distance in number of bases of sequence between features are indicated in black above the *FLT3* gene model. (B) Observed haplotypes resolved by cloning and sequencing of these regions are depicted.



A. *FLT3* intron 1 indels identified by WGS

B. *FLT3* haplotypes determined by cloning and sequencing

Figure S41. Protein expression (immunohistochemistry) of FLT3 in ALL1 and three reference tumors
5 um paraffin sections from formalin-fixed bone marrow core biopsies were stained with rabbit anti-CD135/FLT3 (Acris Antibodies, San Diego, USA, clone AP21030PU-N) according to standard protocols. AML samples with varying levels of *FLT3* as determined by RNA-seq analysis [45] were first used to determine the specificity of the antibody. Images were acquired at 600x magnification on an Olympus BX60 microscope using an Infinity 3 Lumenera camera. (A) An AML with low FLT3 expression (UPN 884262; FPKM = 3.8). (B) An AML with high FLT3 expression (UPN 972783; FPKM = 65.5). (C) A second AML with high FLT3 expression (UPN 923966; FPKM = 106.9). (D) The second relapse tumor sample from ALL1 demonstrating high FLT3 expression and strong membranous reactivity (FPKM = 108.0).


A. Low expression control
B. High expression control
C. High expression control
D. ALL1

AUTHOR CONTRIBUTIONS

M.G. and O.L.G. led the data analysis. M.G., O.L.G., K.K., Z.L.S., A.R., A.H.W., K.M.C., R.L., J.H., J.Z., N.C.S., C.A.Maher, and B.J.A. performed data analysis, and prepared figures and tables. M.G., O.L.G., K.K., Z.L.S., A.R., A.H.W., K.M.C., R.L., J.H., J.Z., B.J.A., D.E.L., S.M.S., J.R.W., J.M.E., and C.A.Miller wrote code and developed custom software for these analyses. M.J.C, J.M.K, T.L.L, S.H., C.F., S.O., V.M., and S.M. developed laboratory methods and performed laboratory experiments. J.M.K., S.H., J.E.P., M.J.W., D.C.L., T.A.G., P.W., S.K., J.F.D., and T.J.L. contributed to clinical interpretation, patient management, procurement of patient samples, and nucleic acid preparation. S.K. performed cytogenetics. P.W. and T.J.L. wrote the clinical case history. M.G., O.L.G, D.E.L., R.F., J.F.D., E.R.M., R.K.W. and T.J.L. developed the project concept and experimental design. M.G., O.L.G. and T.J.L. wrote the manuscript with input from K.K., Z.L.S., A.R., A.H.W., and E.R.M.

REFERENCES

[1] Griffith M, Miller CA, Griffith OL, et al. Optimizing cancer genome sequencing and analysis. Cell Syst. 2015;1:210-223.

[2] Griffith M, Griffith OL, Smith SM, et al. Genome Modeling System: A Knowledge Management Platform for Genomics. PLoS Comput Biol. 2015;11:e1004274.

[3] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997. 2013.

[4] Pengelly RJ, Gibson J, Andreoletti G, Collins A, Mattocks CJ, Ennis S. A SNP profiling panel for sample tracking in whole-exome sequencing studies. Genome Med. 2013;5:89.

[5] Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-2079.

[6] Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568-576.

[7] McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26:2069-2070.

[8] Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol. 2013;9:e1003153.

[9] Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4:177-183.

[10] Kanchi KL, Johnson KJ, Lu C, et al. Integrated analysis of germline and somatic variants in ovarian cancer. Nat Commun. 2014;5:3156.

[11] Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28:311-317.

[12] Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28:1811-1817.

[13] Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213-219.

[14] Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011;39:D945-950.

[15] Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308-311.

[16] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009;25:2865-2871.

[17] Cunningham F, Amode MR, Barrell D, et al. Ensembl 2015. Nucleic Acids Res. 2015;43:D662-669.

[18] Klambauer G, Schwarzbauer K, Mayr A, et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012;40:e69.

[19] Chen K, Chen L, Fan X, Wallis J, Ding L, Weinstock G. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. Genome Res. 2014;24:310-317.

[20] Spies N, Zook JM, Salit M, Sidow A. svviz: a read viewer for validating structural variants. Bioinformatics. 2015.
[21] Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178-192.
[22] Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. PLoS Comput Biol. 2015;11:e1004393.
[23] Dodt M, Roehr JT, Ahmed R, Dieterich C. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology (Basel). 2012;1:895-905.
[24] Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7:562-578.
[25] Lassmann T, Hayashizaki Y, Daub CO. SAMStat: monitoring biases in next generation sequencing data. Bioinformatics. 2011;27:130-131.
[26] Iyer MK, Chinnaiyan AM, Maher CA. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics. 2011;27:2903-2904.
[27] Zhang J, White NM, Schmidt HK, et al. INTEGRATE: gene fusion discovery using whole genome and transcriptome data. Genome Res. 2016;26:108-118.
[28] Griffith M, Griffith OL, Coffman AC, et al. DGIdb: mining the druggable genome. Nat Methods. 2013;10:1209-1210.
[29] Spencer DH, Young MA, Lamprecht TL, et al. Epigenomic analysis of the HOX gene loci reveals mechanisms that may control canonical expression patterns in AML and normal hematopoietic cells. Leukemia. 2015;29:1279-1289.
[30] Kang H, Chen IM, Wilson CS, et al. Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. Blood. 2010;115:1394-1405.
[31] Miller CA, White BS, Dees ND, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol. 2014;10:e1003665.
[32] Nielsen M, Lundegaard C, Worning P, et al. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. 2003;12:1007-1017.
[33] Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res. 2008;36:W509-512.
[34] Fan X, Abbott TE, Larson D, Chen K. BreakDancer - Identification of Genomic Structural Variation from Paired-End Read Mapping. Curr Protoc Bioinformatics. 2014;2014.
[35] Nickerson DA, Tobe VO, Taylor SL. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. 1997;25:2745-2751.
[36] Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998;8:186-194.
[37] Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. Genome Res. 1998;8:195-202.
[38] Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008;456:66-72.
[39] Shah SP, Morin RD, Khattra J, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature. 2009;461:809-813.
[40] Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113-1120.
[41] International Cancer Genome C, Hudson TJ, Anderson W, et al. International network of cancer genome projects. Nature. 2010;464:993-998.
[42] Zhang J, Ding L, Holmfeldt L, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature. 2012;481:157-163.
[43] Welch JS, Ley TJ, Link DC, et al. The origin and evolution of mutations in acute myeloid leukemia. Cell. 2012;150:264-278.
[44] Holmfeldt L, Wei L, Diaz-Flores E, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013;45:242-252.
[45] Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013;368:2059-2074.

[46] Ma X, Edmonson M, Yergeau D, et al. Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. Nat Commun. 2015;6:6604.
[47] Mullighan CG. The molecular genetic makeup of acute lymphoblastic leukemia. Hematology Am Soc Hematol Educ Program. 2012;2012:389-396.
[48] Gocho Y, Kiyokawa N, Ichikawa H, et al. A novel recurrent EP300-ZNF384 gene fusion in B-cell precursor acute lymphoblastic leukemia. Leukemia. 2015;29:2445-2448.
[49] Martini A, La Starza R, Janssen H, et al. Recurrent rearrangement of the Ewing's sarcoma gene, EWSR1, or its homologue, TAF15, with the transcription factor CIZ/NMP4 in acute leukemia. Cancer Res. 2002;62:5408-5412.
[50] La Starza R, Aventin A, Crescenzi B, et al. CIZ gene rearrangements in acute leukemia: report of a diagnostic FISH assay and clinical features of nine patients. Leukemia. 2005;19:1696-1699.
[51] Zhong CH, Prima V, Liang X, et al. E2A-ZNF384 and NOL1-E2A fusion created by a cryptic t(12;19)(p13.3; p13.3) in acute leukemia. Leukemia. 2008;22:723-729.
[52] Nyquist KB, Thorsen J, Zeller B, et al. Identification of the TAF15-ZNF384 fusion gene in two new cases of acute lymphoblastic leukemia with a t(12;17)(p13;q12). Cancer Genet. 2011;204:147-152.
[53] 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. 2002;99:6567-6572.
[54] Roberts KG, Morin RD, Zhang J, et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. Cancer Cell. 2012;22:153-166.
[55] Roberts KG, Mullighan CG. Genomics in acute lymphoblastic leukaemia: insights and treatment implications. Nat Rev Clin Oncol. 2015;12:344-357.
[56] 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;116:4874-4884.
[57] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5:557-572.