Outlier expression of isoforms by targeted RNA sequencing as clinical markers of genomic variants in B lymphoblastic leukemia and other tumor types

Harrison K. Tsai[1,2], Tasos Gogakos[3], Va Lip[1], Jonathan Tsai[2], Yen-Der Li[4], Adam Fisch[3], Jonathan Weiss[1], Leslie Grimmett[1], Thai Hoa Tran[5,6], Maxime Caron[6], Sylvie Langlois[6], Daniel Sinnett[5,6], Yana Pikman[7,8], Annette S. Kim[2], Valentina Nardi[3], Lewis B. Silverman[7,8], Marian H. Harris[1]

[1]*Department of Pathology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA*
[2]*Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
[3]*Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA*
[4]*Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA*
[5]*Division of Pediatric Hematology-Oncology, Centre Hospitalier Universitaire Sainte-Justine, Montreal, QC, Canada*
[6]*Immune Diseases and Cancers Axis, Centre Hospitalier Universitaire Sainte-Justine Research Center, Montreal, QC, Canada*
[7]*Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA*
[8]*Division of Hematology/Oncololgy, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA*

Corresponding Authors:

Harrison K. Tsai, MD, PhD
Boston Children's Hospital
300 Longwood Ave
Boston, MA 02215
E-mail: Harrison.Tsai@childrens.harvard.edu

Marian H. Harris, MD, PhD
Boston Children's Hospital
300 Longwood Ave
Boston, MA 02215
E-mail: Marian.Harris@childrens.harvard.edu

Word count (abstract): 247, Word count (text): 4139

Figures: 7, Tables: 1, References: 31

Supp Figs: 6

Abstract

Recognition of aberrant gene isoforms indicative of underlying DNA events can impact molecular classification and risk stratification of B lymphoblastic leukemia (B-ALL). Aberrant *ERG* isoforms have been proposed as markers of the favorable-risk *DUX4*-rearranged (*DUX4*r) subtype while deletion-mediated *IKZF1* isoforms are associated with adverse prognosis in non-*DUX4*r B-ALL. The high-risk *IKZF1*^{plus} signature depends on gene deletions including *PAX5* while intragenic *PAX5* amplifications (PAX5amp) are recurrent in the provisional B-ALL with *PAX5*-alteration subtype. In this study, we screened for outlier expression of isoforms within targeted RNA sequencing assays designed for fusions. Outlier analysis of known and novel *IKZF1*, *ERG*, and *PAX5* isoforms was 97.0% (32/33), 90% (9/10), and 100% (6/6) sensitive and 97.8% (226/231), 100% (35/35), and 88.5% (23/26) specific for *IKZF1* intragenic or 3' deletions, *DUX4*r, and *PAX5* intragenic deletions respectively, where false positives were favored to represent low-level deletions below the limit of DNA-based detection. Outlier analysis also identified putative PAX5amp cases and revealed partial tandem duplication (PTD) spanning *IKZF1* N159Y in the B-ALL with mutated N159Y subtype. To demonstrate utility in other tumor types, outlier analysis was 100% (9/9) sensitive and 100% (255/255) specific for *KMT2A*-PTD in hematologic samples and 100% (7/7) sensitive and 100% (79/79) specific for *FGFR1* tyrosine kinase domain duplication in brain tumors. These findings support the use of aberrant isoform analysis in targeted RNA sequencing data as a robust strategy for the detection of clinically significant DNA events.

Introduction

The versatile applications of gene fusions associated with chromosomal rearrangements towards diagnosis, prognostication, treatment, and surveillance of disease has motivated a growing use of targeted RNA sequencing in clinical laboratories. Gene isoforms, characterized by exon skipping, out-of-order exons, cryptic splice sites, or other aberrations, may similarly reflect deletions, duplications, spice variants, or other underlying DNA alterations of clinical importance, though the use of targeted RNA sequencing for purpose of their identification is less well established. Molecular classification and risk stratification of pediatric B lymphoblastic leukemia (B-ALL) may particularly benefit from recognition of aberrant isoforms. *ERG* isoforms induced by *DUX4* binding (ERGaltA/B) or mediated by intragenic *ERG* deletions have been proposed as markers of the favorable-risk *DUX4* rearranged (*DUX4*r) subtype, which occurs in 4-14% of pediatric B-ALL but has typically required whole genome sequencing or whole transcriptome RNA sequencing (wtRNAseq) for detection[1-3]. In non-*DUX4r* B-ALL, dominant negative and loss-of-function *IKZF1* isoforms arising from intragenic *IKZF1* deletions have been associated with a worse prognosis while *PAX5* isoforms with out-of-order exons have been associated with intragenic *PAX5* amplification (PAX5amp) within the emerging *PAX5*-altered subtype (PAX5alt)[4-6]. Patterns of isoform co-occurrence also carry clinical significance, highlighted by the high-risk signature *IKZF1*^{plus} defined by various co-deletions[7-9]. Several clinically important DNA variants generate aberrant RNA isoforms outside of B-ALL, including *KMT2A*-PTD in myeloid malignancies, *FGFR1* tyrosine kinase domain duplication (*FGFR1*-TKDD) in low grade gliomas, and many others[10-11].

The isoforms described above each harbor an isoform-specific junction, however RNA transcripts from sources other than DNA aberrations can a priori contain identical junctions. Native alternative splicing and alternative transcriptional initiation are theoretically capable of producing exon-skipping isoforms and ERGalt transcripts, while circular RNAs (circRNAs) might produce the out-of-order exons of PAX5amp or *KMT2A*-PTD. Since these other sources generally produce RNA at low levels, characterization of their empiric background distribution is important to optimize sensitivity and specificity of aberrant isoforms as NGS markers of DNA events. In this study, we apply systematic outlier analysis to targeted RNA next generation sequencing (NGS) data based on anchored multiplex PCR (AMP), a popular enrichment method for detecting fusions[12]. We validate outlier expression of select isoforms as markers of underlying genomic variants and more generally describe informatic strategies for evaluating intragenic variants within a clinical practice.

Methods

Sample selection for targeted RNA NGS: Total nucleic acid extracted from bone marrow, peripheral blood, cell lines, or formalin-fixed paraffin-embedded tissue of other disease sites was tested at Boston Children's Hospital by 1 of 2 clinically validated targeted NGS panels based on AMP (ArcherDx, Boulder, CO): i) ArcherDx FusionPlex

Heme v2 assay (FPH) performed clinically, for research, or for validation on hematologic samples from 2014-2022 (heme cohort; n = 407) comprising predominantly pediatric cases but also including known *KMT2A*-PTD adult samples and the EOL-1 cell line or ii) Custom Solid and Brain Tumor Fusion Panel (SBT) performed clinically on pediatric brain samples (n=86) and solid tumor samples (n=95) from 2020-2021, each followed by 2 x 151 bp paired-end Illumina sequencing and bioinformatic processing by Archer Analysis v6.2.7 (AA) under default parameters. NGS samples yielding under 25,000 unique RNA reads were excluded from the above cohorts. The study was conducted in accordance with the Helsinki Declaration and with approval of Institutional Review Boards.

Isoform junction notation, split-read analysis, and outlier criteria: We used the notation *IKZF1* e3e8 to denote the junction connecting the 3' end of *IKZF1* exon 3 to the 5' end of exon 8, and similarly for other genes and exact junctions. For each gene of interest, paired reads were extracted from binary alignment map files if they aligned within 1 Mb of the gene locus and involved both a primary and a supplementary alignment for at least one read of the pair (i.e. a split-read). Primary and supplementary alignments were parsed to determine the genomic coordinates (chrA:x0 and chrA:x1) indicating a breakpoint via soft-clipping or hard-clipping. In this manner, split-reads were associated with isoform junctions of the form "chrA:x0_x1_overlap_n", where overlap was the amount of microhomology (n>0) or unaligned nucleotides (n<0) around the breakpoints. Paired reads assigned to each isoform junction were counted and sorted further by originating gene specific primer (GSP2), which was inferred from the alignment start coordinates of anchored second mate reads. Based on these paired-read assignments, expressed variant allele fraction (VAF) of an isoform junction was calculated relative to all isoform junctions sharing a breakpoint within 10 bp but restricted to split-reads originating from a GSP2 targeting the appropriate side of the shared breakpoint. Normalized expression of an isoform junction was defined by dividing split-read counts by total unique RNA reads. Isoform junctions were annotated by comparing against exon coordinates of transcripts from NCBI Refseq and Gencode to test equivalence with exact exon-exon junctions; otherwise, an isoform junction was considered novel and potentially represented use of a cryptic splice site. Isoform junctions with both breakpoints contained within a single intron of all known transcripts were predicted to have minimal effect and were filtered out. For remaining isoform junctions, outlier expression was defined as satisfying both VAF $\geq$ max(0.01, 5[90% quantile VAF expression]) and normalized expression $\geq$ 0.0001.

Whole-transcriptome RNA-sequencing for *DUX4*r status: *DUX4*r status was available for select cases (n=45) from the heme cohort based on clinical reports from whole transcriptome RNA-sequencing tests performed at Centre Hospitalier Universitaire Sainte Justine[13].

DNA-based testing: Most (n=264) of the heme cohort and all pediatric brain samples underwent concurrent clinical testing by 1 of 2 DNA-based targeted NGS panels performed clinically at Brigham and Women's Hospital: (i) Rapid Heme Panel version 3 (RHP) based on NEBNext Direct capture/amplicon hybrid chemistry (NEB; New England BioLabs, Ipswich, MA)[14] or (ii) Oncopanel (OP) based on Agilent SureSelect

hybridization capture probes (Agilent Technologies, Santa Clara, CA)[15]. RHP and OP were clinically validated to detect copy number changes of *IKZF1*, *ERG*, and *KMT2A*, and OP was also validated to detect structural rearrangements of *FGFR1.* RHP did not routinely report involved exons, thus deletion status of *IKZF1* and *ERG* on the level of individual targeted exons was re-interpreted independently by 3 molecular pathologists (HKT, MHH, TSG) based on supplementary analysis of read count data (modified BatchCNV)[16]. Copy number of *PAX5*, *IKZF1*, and *ERG* were additionally available for a limited subset (n=32) of the heme cohort based on (i) clinical reports from cytogenetic microarray testing (CGX Onco Array; PerkinElmer, Waltham, MA) performed at Centre Hospitalier Universitaire Sainte Justine (n=7) or (ii) multiplex ligation-dependent probe amplification (MLPA) (Probemix P335; MRC Holland, Amsterdam, the Netherlands) performed for this study to confirm findings or resolve discrepancies (n=29).

Gene Expression by targeted RNA NGS: The normalized expression of a gene was calculated by summing unique RNA reads derived from all GSP2s targeting the gene and dividing by total number of unique RNA reads in the sample. Read counts were provided by AA output. No adjustments were made for the number of GSP2s, length of a gene, or batch effects. t-distributed stochastic neighbor embedding (t-SNE) plots were generated from log-transformed normalized expression using the R package Rtsne with perplexity set to 5 and default settings for other parameters. Annotations were based on available data from karyotype, FISH, FPH, RHP, OP, wtRNAseq, and low density arrays (LDA).

Results

Outlier analysis identifies known and novel *IKZF1*, *ERG*, and *PAX5* isoform junctions with high specificity for B-ALL.

To investigate utility of aberrant isoforms, we screened for outlier expression of *IKZF1*, *ERG*, and *PAX5* isoforms in the heme cohort. By using an annotation-independent approach to maximize isoform recognition, we identified novel recurrent junctions (Fig 1) connecting *IKZF1* exons 2 or 3 to intergenic breakpoints (chr7:50498139, chr7:50505862, chr7:50525921) adjacent to putative cryptic splice acceptor sites approximately 25-50 kb downstream of *IKZF1*, thereby compatible with previously characterized DNA breakpoints associated with deletions involving exons 3-8 or 4-8[17,18]. Our approach similarly enabled detection of known pathogenic cryptic isoforms such as ERGalt despite their absence from annotation databases. Overall we identified outlier expression of 57 isoform junctions, with high specificity for a B-ALL component (96.5%; 193/200) and modest sensitivity for B-ALL (41.3%; 85/206) or MPAL with B-ALL (100%; 1/1) (SFig 1). A few non-specific (non B-ALL) cases narrowly exceeded outlier thresholds of common alternative splice junctions (*ERG* e6e8, *PAX5* e6e9), and were favored to reflect suboptimally low thresholds.

IKZF1 exon-skipping isoforms at outlier expression levels are sensitive and specific markers for underlying intragenic and 3' deletions.

We compared RNA outlier expression status of *IKZF1* exon-skipping isoforms to DNA deletion status of the skipped exons (Figure 2, Table 1). We found outlier expression of the novel junctions e3:g.50505862 or e2:g.50505862 to be 100% sensitive and 100% specific for deletions of exons 4-8 (n=5) or exons 3-8 (n=2) respectively. Similarly, outlier expression of e1e4, e1e5 (without e1e4), e2e4, or e1e3 was 100% sensitive and 100% specific for uncommon deletions involving exons 2-3 (n=2), exons 2-4 (n=1), exon 3 (n=2), or exon 2 (n=1) respectively. Outlier expression of e1e8 was 100% sensitive (5/5) but 98.8% (256/259) specific for deletion of exons 2-7. Two of the false positives were associated with slightly different deletions (exons 1-7 and 2-8 instead of 2-7) incompatible as single events with formation of e1e8, thus raising the possibility of mischaracterized deletion boundaries or multiple deletions on separate alleles (e2-e7 on one allele combined with e1 or e8 on the other allele). Finally, outlier expression of e3e8, corresponding to the dominant negative isoform Ik-6, was the most prevalent aberration and was 92.9% (13/14) sensitive and 98.4% (246/250) specific for deletion of exons 4-7. The lone false negative had borderline RNA quality with an e3e8 VAF (0.93%) narrowly missing the outlier threshold (1.10%). The 4 false positives were favored to represent deletions of e4-e7 below the limit of detection of DNA-based assessment, given that RNA-based sensitivity for aberrant e3e8 was enhanced in 2 ways compared to DNA tests. First, background noise of e3e8 from presumed alternative splicing was minimal, resulting in a low outlier threshold (VAF 1.10%) well below the limit of detection by standard DNA testing for copy number events (typically VAF $\geq$ 10-20%). Second, RNA-based VAF of e3e8 was usually greater than DNA-based VAF of exon 4-7 deletions (p=0.00006 by paired t-test), empirically strengthening signal to noise and raising the possibility of allelic overexpression (Fig 3).

We also explored whether whole gene or 5' deletions of *IKZF1* might be detectable through haploinsufficient *IKZF1* gene expression but found no significant differences between B-ALL cases with shallow *IKZF1* deletions (p=0.36) versus non-deleted B-ALL (defined as having neither deletions nor outlier isoform status), although a trend appeared for deep *IKZF1* deletions (p=0.06) (SFig 2). Rather, we only found differential *IKZF1* gene expression in e3e8 outliers (p=0.0007; up-regulation) versus non-deleted B-ALL, again suggestive of allelic overexpression.

B-ALL with *IKZF1* N159Y have recurrent *IKZF1*-PTD of exon 5 duplicating N159Y.

We compared outlier status to molecular characteristics (SFig 1). In 2 of 2 B-ALL cases harboring the *IKZF1* N159Y mutation (a proposed emerging subtype with a distinct gene expression profile[5]), we identified outlier expression of *IKZF1* e5e5 (VAFs 0.443-0.517) compared to zero expression in virtually all remaining cases (396/397), except for 1 case with 1 e5e5 read (VAF 0.006). *IKZF1* exon 5 was not targeted by FPH, arguing against circRNA of *IKZF1* exon 5 as a source of e5e5 reads. Rather, we found e5e5 reads in both cases derived from all 5 *IKZF1* primers (targeting forward strands of exons

1-3 and reverse strands of exons 7-8), consistent with mRNA transcripts extending from exon 1 to exon 8 with PTD of exon 5. Since *IKZF1* N159Y is located within exon 5, we asked whether N159Y or PTD occurred first or independently in clonal evolution. By identifying e5e5 reads with N159Y on either side of the e5e5 junction, we deduced PTD as the secondary event in both cases (SFig 3). RHP did not report copy number gains of exon 5, thus we manually reviewed copy number data and raw RHP alignments over *IKZF1*, revealing small partial tandem duplications (PTD) of sizes 344 bp and 396 bp spanning *IKZF1* exon 5 and duplicating N159Y (Fig 4). Given this unexpected finding, we screened the entire historical RHP cohort. We did not find PTD of exon 5 in multiple cases with N159S/T, and there were no further instances of N159Y. However, among AML cases without N159 variants, we identified very rare in-frame ITDs within exon 5 spanning the N159 codon.

In contrast to *IKZF1* e5e5 and its association with PTD of exon 5, our screening approach yielded 2 samples (1 B-ALL and 1 eosinophilia) with outlier expression of *IKZF1* e3e2 minimally exceeding thresholds, where the pattern of originating primers of e3e2 reads was consistent with circRNA of *IKZF1* e2-e3 and not PTD of exons 2-3. Namely, e3e2 reads were derived almost exclusively from primers targeting forward strands of exons 2-3 whereas significant fractions of e3e4 reads were derived from primers targeting each of the exons 1-3.

Outlier expression of either ERGalt A or *ERG* isoforms is highly sensitive and specific for *DUX4*-rearranged B-ALL.

We investigated the utility of outlier *ERG* isoforms as markers for *DUX4*r B-ALL. We found that outlier expression of either (i) ERGaltA (threshold of 11.0% VAF) or (ii) *ERG* e4e10 or e4e12 (thresholds of 1%) was 90% (9/10) sensitive and 100% (35/35) specific for *DUX4*r status determined by wtRNAseq (Fig 5). Both criteria (i) and (ii) contributed to optimizing performance since their sensitivities as standalone markers were 80% and 50% respectively. We identified outlier expression in 8 additional cases without wtRNAseq data; all were B-ALLs lacking an established molecular subtype, including 2 with intragenic *ERG* deletions by DNA NGS and 1 with a DUX4/ERG signature by LDA screening. Among the overall cohort, cases with deletions of *ERG* exons 5-9 or 5-11 were nearly always (86%; 6/7) associated with outlier e4e10 or e4e12 expression, where the case lacking isoform expression was complicated by borderline RNA quality (same false-negative case with low-level outlier *IKZF1* e3e8); by contrast, cases with deletions of *ERG* exons 4-7 (n=2) or exon 4 (n=1) had zero expression of corresponding isoforms e3e8 or e3e5.

We observed high *PDGFRA* gene expression in the lone false negative case not identified by outlier *ERG* analysis. *PDGFRA* overexpression has been associated with *ERG* deletions and binding of DUX4 to the *PDGFRA* transcription start site in *DUX4*r[1,19], however we also observed high expression in a *DUX4*r-negative B-ALL in our cohort, limiting its specificity. Given the known distinct gene expression profile of *DUX*4r B-ALL, we analyzed for differential gene expression among the 100 genes targeted by FPH,

yielding 8 genes with significant overexpression in *DUX4*r-positive versus *DUX4*r-negative B-ALL. These genes demonstrated at most 30% sensitivity to achieve 100% specificity for *DUX4*r, thus ERGaltA had the top performance as a single positive marker for *DUX4*r (80% sensitivity, 100% specificity). Similarly, tSNE relative to FPH gene expression data generated a cluster consisting predominantly but not exclusively of *DUX4*r (SFig 4).

PAX5amp is associated with *PAX5* out-of-order isoforms at high VAFs.

We screened *PAX5* isoforms harboring out-of-order exons at outlier levels for potential PAX5amp candidates. We observed robust background expression of e5e2 across B-ALL cases, compatible with known circRNA involving *PAX5* exons 2-5 and resulting in a relatively high outlier threshold. We identified 4 cases (P1-P4) with clear outlier expression of e5e2 (67.7% VAF), e5e2 (21.8%), joint e7e2 (55.0%) and e5e2 (20.5%), and e7e7 (21.5%) respectively (Fig 6A). Using MLPA, we confirmed parsimonious underlying *PAX5* copy number alterations characterized by (P1) high level gain of exons 2-5, (P2) shallow gain of exons 1-5 (compatible with e5e2 isoforms since e1 lacks a splice acceptor to enable e5e1 splicing), (P3) several different copy number levels suggesting multiple gain events involving exons 1-7, 2-7, and 2-5, and (P4) shallow gain of exon 7 (Fig 6B-E). The cases P1 and P3, with high RNA VAFs above 50% and high level copy number gains, lacked an alternative molecular subtype, arguing for the emerging PAX5amp subtype, whereas P2 harbored an ABL class fusion (*PDGFRB*), which has been previously described in this context[5], and P4 had a clonal *PCF3-TBX1* rearrangement (SFig 5). Two additional cases narrowly exceeded outlier thresholds of e7e2 and e4e2 and were favored to represent circRNAs and suboptimal thresholds; the e4e2 case was associated with a *PAX5* fusion with an exon 4 breakpoint, suggesting altered circRNA formation in the context of rearrangement (Fig 6A, SFig 5). Limited MLPA and microarray testing of cases without outlier expression of out-of-order *PAX5* isoform junctions identified no other instances of intragenic *PAX5* gains (n=28).

***PAX5* exon-skipping isoforms at outlier levels are associated with intragenic deletions and facilitate specific but not sensitive detection of the *IKZF1*^{plus} signature.**

We next screened exon-skipping *PAX5* isoforms for potential intragenic deletions (SFig 5-6). We identified outlier expression across 30 samples involving 11 junctions (e1e6, e1e7, e1e8, e1e9, e5e7, e5e8, e5e9, e5e10, e6e8, e7e9, e8e10). Outlier e5e8, e5e9, e7e9, or e8e10 frequently co-occurred with *CRLF2* rearrangements, while outlier e1e7, e1e8, or e1e9 frequently co-occurred with *ETV6-RUNX1*, as has been previously described (SFig 5)[20]. We observed single exon-skipping isoforms (e7e9, e8e10) at high outlier VAFs (>85%) favoring loss of heterozygosity in 3 cases, and outlier co-expression of multiple isoforms similarly raising the possibility of biallelic dysfunction in 2 other cases (Q4-Q5). We compared outlier expression to intragenic deletions in a

limited cohort (n=32) and found outlier expression of e1e7, e7e9, e5e8, e5e9, or e5e10 to be 100% (6/6) sensitive (cases Q1-Q6) and 88.5% (23/26) specific for intragenic deletions. The 3 false positives (Q1-Q3) all had low outlier VAFs (< 20%) and were suspected to represent deletions below the limit of detection of DNA testing, similar to *IKZF1* findings. The 2 cases (Q4-Q5) with outlier co-expression of multiple isoforms demonstrated complicated copy number profiles implying multiple intragenic deletion events.

We evaluated for the B-ALL high-risk *IKZF1*^{plus} signature, defined by *IKZF1* deletion combined with one or more deletions of *PAX5*, *CKDN2A*, *CDKN2B*, and/or the PAR1 region in the absence of *ERG* deletion[7]. We used outlier expression of *IKZF1*, *PAX5*, and *ERG* exon-skipping isoforms as makers of focal deletions, and used the fusion *P2RY8-CRLF2*, whose detection was clinically validated under FPH, as a proxy for PAR1 deletion. FPH accordingly identified 10 B-ALL cases with focal *IKZF1* deletions satisfying *IKZF1*^{plus}, composed of 4 with inferred *PAX5* intragenic deletions alone, 4 with *P2RY8-CRLF2* alone, and 2 with both. Sensitivity for *IKZF1*^{plus} was limited by inability to detect broad *IKZF1*, *PAX5*, and *CDKN2A* deletions and a lack of *CDKN2B* targets.

Outlier expression of *KMT2A* e8e2, e10e2, e8e4, or e6e3 is sensitive and specific for the corresponding *KMT2A*-PTD at initial diagnosis.

We tested the ability of outlier isoform expression to detect *KMT2A*-PTD as determined by clinically validated orthogonal testing, where the known *KMT2A*-PTD cases (1 pediatric, 7 adult, and the cell line EOL1) corresponded to the isoforms e8e2 (5), e10e2 (2), e8e4 (1), and e6e3 (1). Positive (n=9) and negative (n=255) *KMT2A*-PTD cases were completely separated by outlier expression thresholds (Figure 7A). Rare non-PTD cases narrowly exceeded VAF cutoffs for e8e2 but remained below the normalized expression cutoff (Figure 7B); these occurred in *KMT2A*-rearranged leukemias involving exon 8, consistent with altered circRNA formation as described previously[21]. The *KMT2A* junction e8e4 had the highest background expression, presumably derived from the known circRNA species hsa_circ_0095112[22]. By contrast, *KMT2A* junction e6e3 was expressed in only 2 non-PTD cases at 1 split-read.

Split-read analysis of *FGFR1* breakpoints spanning the kinase domain is sensitive and specific for *FGFR1*-TKDD.

We applied outlier analysis to *FGFR1*-TKDD. In brain tumor samples tested by SBT, the detection of split-reads with a 5' breakpoint within exon 18 and a 3' breakpoint within exon 9 or intron 9 was 100% (7/7) sensitive and 100% (79/79) specific for detection of *FGFR1-TKDD* relative to Oncopanel, where the positive cases were all gliomas. RNA mutant breakpoints under SBT agreed with genomic DNA breakpoints under Oncopanel in all positive samples. No solid tumor samples (non-brain) demonstrated split-reads characteristic of *FGFR1*-TKDD. Given the fact that breakpoints in exon 18 were

intraexonic and not adjacent to potential cryptic splice sites, low level expression from alternative sources such as circRNA was never present.

Discussion.

Contemporary B-ALL risk stratification utilizes *IKZF1* deletion status to improve treatment outcomes[23]. In FPH, we found outlier expression of known and novel *IKZF1* exon-skipping isoforms to be effective markers of intragenic and 3' deletions with 97.0% (32/33) sensitivity and 97.8% (226/231) specificity overall. Moreover, the false negative was associated with suboptimal RNA quality, and the false positives were either suspected to represent subclonal deletions below the limit of detection of our DNA-based assays or were due to minor discrepancies in deletion boundaries. Targeted RNA sequencing may accordingly be more reliable than wtRNAseq, which has successfully detected deletions of exons 4-7 through the e3e8 isoform but has been less successful with other deletions including exons 4-8[13,24]. In FPH, accurate quantification and optimal distinction from background noise may have benefited from direct targeting of relevant junctions, deeper coverage, and use of unique molecular identifiers. Alternatively, it is possible that wtRNAseq performance may improve from different informatic approaches, including the outlier methods used here or the recently developed MINTIE algorithm[25]. However, the MINTIE study reported a limited number of cases (n=7) with *IKZF1* RNA variants and may not have had orthogonal DNA data to assess performance. Thus, head-to-head comparisons with uniform informatics of targeted RNA sequencing and wtRNAseq may be informative.

In FPH, RNA VAFs of e3e8 were significantly greater than DNA VAFs of exon 4-7 deletions, suggesting allelic overexpression and empirically improving limit of detection regardless of whether mediated biologically or by assay bias; this behavior might also partially explain the successful use of e3e8 in wtRNAseq studies. More generally, aberrant isoforms across genes frequently demonstrated minimal background noise and defaulted to 1% outlier thresholds, potentially enabling subclonal detection of underlying DNA events well below the limit of detection of standard DNA-based assays[26,27]. Copy number changes involving a single exon present another challenging category for DNA-based assays, since their signal may be identical to noise, whereas outlier isoform analysis can more reliably ensure the high specificity required for clinical diagnostics. We identified unrecognized recurrent PTDs of *IKZF1* exon 5 duplicating N159Y on both the RNA and DNA levels in the B-ALL subtype with mutated N159Y. RNA e5e5 reads were similarly detected by MINTIE in a B-ALL with the gene expression signature of *IKZF1* N159Y, although mutational status was not described[25]. The significance of PTD and its contribution to the distinct gene expression profile associated with N159Y will require further study.

Robust isoform detection based on systematic outlier analysis importantly facilitates recognition of new genetically defined subtypes of B-ALL specified by the International Consensus Classification and the 5th edition of the WHO, thereby complementing fusion detection to enable more complete molecular classification within a single

targeted RNA sequencing assay[28]. We found outlier expression of *ERG* isoforms to be 90% (9/10) sensitive and 100% (35/35) specific for the relatively frequent *DUX4*r subtype, which has been underrecognized historically since the rearrangements involve cytogenetically cryptic small insertions that are also difficult to molecularly target due to the location of *DUX4* within a D4Z4 macrosatellite repeat array[30]. Performance of these *ERG* isoform markers has been less strong in wtRNAseq[1,2], similar to *IKZF1*. False negatives in FPH may be rescued by gene expression screening, however *DUX4*r candidates identified in this way should be confirmed by orthogonal methods due to imperfect specificity of gene expression analyses. FPH may also be capable of detecting the complete spectrum of *PAX5* variants (P80R, rearrangements, and intragenic amplifications) defining the B-ALL with mutated *PAX5* P80R subtype and the provisional B-ALL with *PAX5* alteration subtype. PAX5amp e5e2 transcripts must be distinguished from circRNA of *PAX5* exons 2-5, which is upregulated in B-cells and captured by assays lacking a polyA mRNA enrichment step[31,32]. In our cohort, e5e2 junctions were present in 98.7% of B-ALL samples, highlighting the importance of accurate quantification.

Outlier analysis is most effective for genomic events that yield robust expression of RNA isoforms but may be challenging for poorly expressed genes. Our systematic rule for defining outlier thresholds was also relatively arbitrary, and performance may be improved by selective isoform-specific modifications. Overall however, we find that properly defined outlier analysis of RNA isoforms can be used to identify multiple clinically relevant genomic variants, and hypothesize that this analysis can be extended to additional genomic loci of clinical significance.

Acknowledgements

This work was supported by a St. Baldrick's Foundation Consortium grant, Charles H. Hood Foundation Grant (YP), NIH K08 CA222684 (YP), the clinical research staff of Dana-Farber Cancer Institute (Asmani Adhav, Hannah McLane), the study team of the DFCI ALL Consortium Protocol 16-001, the staff of the Center for Advanced Molecular Diagnostics at Brigham and Women's Hospital, and the staff of the Laboratory for Molecular Pediatric Pathology at Boston Children's Hospital. We also thank the lab of Benjamin Ebert for the gift of EOL-1 cells.

Author Contributions

HKT and MHH designed the study and wrote the manuscript. THT, YP, ASK, VN, LBS, and MHH contributed data. VL, JT, YL, and LG performed experimental work. HKT developed informatic tools. All authors performed data analysis and edited the manuscript.

Disclosure of Conflicts of Interest

The authors declare no competing financial interests.

References

1. Zhang J, McCastlain K, Yoshihara H, Xu B, Chang Y, Churchman ML, et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. Nat Genet 2016;48(12):1481-1489. (PMID: 27776115)
2. Zaliova M, Potuckova E, Hovorkova L, Musilova A, Winkowska L, Fiser K, Stuchly J, Mejstrikova E, Starkova J, Zuna J, Stary J, Trka J. ERG deletions in childhood acute lymphoblastic leukemia with DUX4 rearrangements are mostly polyclonal, prognostically relevant and their detection rate strongly depends on screening method sensitivity. Haematologica 2019;104(7):1407-1416. (PMID: 30630977)
3. Li Z, Lee SHR, Chin WHN, Lu Y, Jiang N, Evelyn Huizi Lim, Coustan-Smith E, Chiew KH, Oh BLZ, Koh GS, Chen Z, Kham SKY, Quah TC, Lin HP, Tan AM, Ariffin H, Yang JJ, Yeoh AE. Distinct clinical characteristics of DUX4- and PAX5-altered childhood B-lymphoblastic leukemia. Blood Adv 2021; 5(23):5226-5238. (PMID: 34547766)
4. Mullighan CG, Su X, Zhang J, Radtke I, Phillips LA, Miller CB, Ma J, Liu W, Cheng C, Schulman BA, Harvey RC, Chen IM, Clifford RJ, Carroll WL, Reaman G, Bowman WP, Devidas M, Gerhard DS, Yang W, Relling MV, Shurtleff SA, Campana D, Borowitz MJ, Pui CH, Smith M, Hunger SP, Willman CL, Downing JR, Children's Oncology Group. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N Engl J Med 2009;360(5):470-80. (PMID: 19129520)
5. Gu Z, Churchman ML, Roberts KG, Moore I, Zhou X, Nakitandwe J, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. Nat Genet 2019;51(2):296-307. (PMID: 30643249)
6. Zaliova M, Stuchly J, Winkowska L, Musilova A, Fiser K, Slamova M, Starkova J, Vaskova M, Hrusak O, Sramkova L, Stary J, Zuna J, Trka J. Genomic landscape of pediatric B-other acute lymphoblastic leukemia in a consecutive European cohort. Haematologica 2019;104(7):1396-1406. (PMID: 30630978)
7. Stanulla M, Dagdan E, Zaliova M, Moricke A, Palmi C, Cazzaniga G, et al. IKZF1 plus Defines a New Minimal Residual Disease-Dependent Very-Poor Prognostic Profile in Pediatric B-Cell Precursor Acute Lymphoblastic Leukemia. J Clin Oncol 2018;36(12):1240-1249. (PMID: 29498923)
8. Clappier E, Auclerc MF, Rapion J, Bakkus M, Caye A, Khemiri A, Giroux C, Hernandez L, Kabongo E, Savola S, Leblanc T, Yakouben K, Plat G, Costa V, Ferster A, Girard S, Fenneteau O, Cayuela JM, Sigaux F, Dastugue N, Suciu S, Benoit Y, Bertrand Y, Soulier J, Cavé H. An intragenic ERG deletion is a marker of an oncogenic subtype of B-cell precursor acute lymphoblastic leukemia with a favorable outcome despite frequent IKZF1 deletions. Leukemia 2014;28(1):70-7. (PMID: 24064621)
9. Zaliova M, Zimmermannova O, Dörge P, Eckert C, Möricke A, Zimmermann M, Stuchly J, Teigler-Schlegel A, Meissner B, Koehler R, Bartram CR, Karawajew L, Rhein P, Zuna J, Schrappe M, Cario G, Stanulla M. ERG deletion is associated with CD2 and attenuates the negative impact of IKZF1 deletion in childhood acute lymphoblastic leukemia. Leukemia 2014;28(1):182-5. (PMID: 24072102)

10. Schichman SA, Caligiuri MA, Gu Y, Strout MP, Canaani E, Bloomfield CD, Croce CM. ALL-1 partial duplication in acute leukemia. Proc Natl Acad Sci U S A. 1994;91(13):6236-9. (PMID: 8016145)
11. Ryall S, Zapotocky M, Fukuoka K, Nobre L, Guerreiro Stucklin A, Bennett J, Siddaway R, Li C, Pajovic S, Arnoldo A, Kowalski PE, Johnson M, Sheth J, Lassaletta A, Tatevossian RG, Orisme W, Qaddoumi I, Surrey LF, Li MM, Waanders AJ, Gilheeney S, Rosenblum M, Bale T, Tsang DS, Laperriere N, Kulkarni A, Ibrahim GM, Drake J, Dirks P, Taylor MD, Rutka JT, Laughlin S, Shroff M, Shago M, Hazrati LN, D'Arcy C, Ramaswamy V, Bartels U, Huang A, Bouffet E, Karajannis MA, Santi M, Ellison DW, Tabori U, Hawkins C. Integrated Molecular and Clinical Analysis of 1,000 Pediatric Low-Grade Gliomas. Cancer Cell. 2020;37(4):569-583. (PMID: 32289278)
12. Zheng Z, Liebers M, Zhelyazkova B, Cao Y, Panditi D, Lynch KD, Chen J, Robinson HE, Shim HS, Chmielecki J, Pao W, Engelman JA, Iafrate AJ, Le LP. Anchored multiplex PCR for targeted next-generation sequencing. Nat Med 2014;20(12):1479-84. (PMID: 25384085)
13. Tran TH, Langlois S, Meloche C, Caron M, Saint-Onge P, Rouette A, Bataille AR, Jimenez-Cortes C, Sontag T, Bittencourt H, Laverdière C, Lavallée VP, Leclerc JM, Cole PD, Gennarini LM, Kahn JM, Kelly KM, Michon B, Santiago R, Stevenson KE, Welch JJG, Schroeder KM, Koch V, Cellot S, Silverman LB, Sinnett D. Whole-transcriptome analysis in acute lymphoblastic leukemia: a report from the DFCI ALL Consortium Protocol 16-001. Blood Adv. 2022 Feb 22;6(4):1329-1341. doi: 10.1182/bloodadvances.2021005634. (PMID: 34933343)
14. Kluk MJ, Lindsley RC, Aster JC, Lindeman NI, Szeto D, Hall D, Kuo FC. Validation and Implementation of a Custom Next-Generation Sequencing Clinical Assay for Hematologic Malignancies. J Mol Diagn. 2016;18(4):507-15. (PMID: 27339098)
15. Garcia EP, Minkovsky A, Jia Y, Ducar MD, Shivdasani O, Gong X, Ligon AH, Sholl LM, Kuo FC, MacConaill LE, Lindeman NI, Dong F. Validation of OncoPanel: A Targeted Next-Generation Sequencing Assay for the Detection of Somatic Variants in Cancer. Arch Pathol Lab Med 2017;141(6):751-758. (PMID: 28557599)
16. Tsai HK, Gibson CJ, Murdock HM, Davineni PK, Harris M, Wang ES, Gondek LP, Kim AS, Nardi V, Lindsley RC. Allelic Complexity of KMT2A Partial Tandem Duplications in Acute Myeloid Leukemia and Myelodysplastic Syndromes. Blood Adv. 2022 May 18:bloodadvances.2022007613. doi: 10.1182/bloodadvances.2022007613. Epub ahead of print. (PMID: 35584376)
17. Caye A, KBeldjord K, Mass-Malo K, Drunat S, Soulier J, Gandemer V, Baruchel A, Bertrand Y, Cave H, Clappier E. Breakpoint-specific multiplex polymerase chain reaction allows the detection of IKZF1 intragenic deletions and minimal residual disease monitoring in B-cell precursor acute lymphoblastic leukemia. Haematologica 2013;98(4):597-601. (PMID: 23065506)
18. Kobitzsch B, Gokbuget N, Schwartz S, Reinhardt R, Bruggemann M, Viardot A, Wasch R, Starck M, Thiel E, Hoelzer D, Burmeister T. Loss-of-function but not dominant-negative intragenic IKZF1 deletions are associated with an adverse

prognosis in adult BCR-ABL-negative acute lymphoblastic leukemia. Haematologica 2017;102(10):1739-1747. (PMID: 28751559)

19. Jerchel IS, Chatzivasileiou D, Hoogkamer AQ, Boer JM, Beverloo HB, Pieters R, den Boer ML. High *PDGFRA* expression does not serve as an effective therapeutic target in *ERG*-deleted B-cell precursor acute lymphoblastic leukemia. Haematologica. 2018;103(2):e73-e77. (PMID: 29170250)
20. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, JDalton JD, Girtman K, Mathew S, Ma J, Pounds SB, Su X, Pui CH, Relling MV, Evans WE, Shurtleff SA, Downing JR. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature 2007;446(7137):758-64. (PMID: 17344859)
21. Dal Molin A, Bresolin S, Gaffo E, Tretti C, Boldrin E, Meyer LH, Guglielmelli P, Vannucchi AM, Te Kronnie G, Bortoluzzi S. CircRNAs Are Here to Stay: A Perspective on the MLL Recombinome. Front Genet 2019;10:88. (PMID: 30815012)
22. Glazar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA 2014;20(11):1666-70. (PMID: 25234927)
23. Yeoh AEJ, Lu Y, Chin WHN, Chiew EKH, Lim EH, Li Z, Kham SKY, Chan YH, Abdullah WA, Lin HP, Chan LL, Lam JCM, Tan PL, Quah TC, Tan AM, Ariffin H. Intensifying Treatment of Childhood B-Lymphoblastic Leukemia With IKZF1 Deletion Reduces Relapse and Improves Overall Survival: Results of Malaysia-Singapore ALL 2010 Study. J Clin Oncol. 2018 Sep 10;36(26):2726-2735. doi: 10.1200/JCO.2018.78.3050. Epub 2018 Jul 25. (PMID: 30044693)
24. Brown LM, Lonsdale A, Zhu A, Davidson NM, Schmidt B, Hawkins A, Wallach E, Martin M, Mechinaud FM, Khaw SL, Bartolo RC, Ludlow LEA, Challis J, Brooks I, Petrovic V, Venn NC, Sutton R, Majewski IJ, Oshlack A, Ekert PG. The application of RNA sequencing for the diagnosis and genomic classification of pediatric acute lymphoblastic leukemia. Blood Adv 2020;4(5):930-942. (PMID: 32150610)
25. Cmero M, Schmidt B, Majewski IJ, Ekert PG, Oshlack A, Davidson NM. MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data. Genome Biol. 2021 Oct 22;22(1):296. (PMID: 34686194)
26. Stanulla M, Cavé H, Moorman AV. IKZF1 deletions in pediatric acute lymphoblastic leukemia: still a poor prognostic marker? Blood. 2020 Jan 23;135(4):252-260. (PMID: 31821407)
27. Benard-Slagter A, Zondervan I, de Groot K, Ghazavi F, Sarhadi V, Van Vlierberghe P, De Moerloose B, Schwab C, Vettenranta K, Harrison CJ, Knuutila S, Schouten J, Lammens T, Savola S. Digital Multiplex Ligation-Dependent Probe Amplification for Detection of Key Copy Number Alterations in T- and B-Cell Lymphoblastic Leukemia. J Mol Diagn. 2017 Sep;19(5):659-672. (PMID: 28736295)
28. Arber DA, Orazi A, Hasserjian RP, Borowitz MJ, Calvo KR, Kvasnicka HM, Wang SA, Bagg A, Barbui T, Branford S, Bueso-Ramos CE, Cortes J, Dal Cin P, DiNardo CD, Dombret H, Duncavage EJ, Ebert BL, Estey E, Facchetti F, Foucar K, Gangat N, Gianelli U, Godley LA, Goekbuget N, Gotlib JR, Hellström-Lindberg E, Hobbs G, Hoffman R, Jabbour EJ, Kiladjian JJ, Larson RA, Le Beau MM, Loh

ML, Löwenberg B, Macintyre EA, Malcovati L, Mullighan CG, Niemeyer CM, Odenike O, Ogawa S, Orfao A, Papaemmanuil E, Passamonti F, Porkka K, Pui CH, Radich JP, Reiter A, Rozman M, Rudelius M, Savona MR, Schiffer C, Schmitt-Graeff A, Shimamura A, Sierra J, Stock W, Stone RM, Tallman MS, Thiele J, Tien HF, Tzankov A, Vannucchi AM, Vyas P, Wei AH, Weinberg OK, Wierzbowska A, Cazzola M, Döhner H, Tefferi A. International Consensus Classification of Myeloid Neoplasms and Acute Leukemia: Integrating Morphological, Clinical, and Genomic Data. Blood. 2022 Jun 29:blood.2022015850. doi: 10.1182/blood.2022015850. Epub ahead of print. (PMID: 35767897)

29. Alaggio R, Amador C, Anagnostopoulos I, Attygalle AD, Araujo IBO, Berti E, Bhagat G, Borges AM, Boyer D, Calaminici M, Chadburn A, Chan JKC, Cheuk W, Chng WJ, Choi JK, Chuang SS, Coupland SE, Czader M, Dave SS, de Jong D, Du MQ, Elenitoba-Johnson KS, Ferry J, Geyer J, Gratzinger D, Guitart J, Gujral S, Harris M, Harrison CJ, Hartmann S, Hochhaus A, Jansen PM, Karube K, Kempf W, Khoury J, Kimura H, Klapper W, Kovach AE, Kumar S, Lazar AJ, Lazzi S, Leoncini L, Leung N, Leventaki V, Li XQ, Lim MS, Liu WP, Louissaint A Jr, Marcogliese A, Medeiros LJ, Michal M, Miranda RN, Mitteldorf C, Montes-Moreno S, Morice W, Nardi V, Naresh KN, Natkunam Y, Ng SB, Oschlies I, Ott G, Parrens M, Pulitzer M, Rajkumar SV, Rawstron AC, Rech K, Rosenwald A, Said J, Sarkozy C, Sayed S, Saygin C, Schuh A, Sewell W, Siebert R, Sohani AR, Tooze R, Traverse-Glehen A, Vega F, Vergier B, Wechalekar AD, Wood B, Xerri L, Xiao W. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Lymphoid Neoplasms. Leukemia. 2022 Jul;36(7):1720-1748. doi: 10.1038/s41375-022-01620-2. Epub 2022 Jun 22. (PMID: 35732829)
30. Lilljebjorn H, Henningsson R, Hyrenius-Wittsten A, Olsson L, Orsmark-Pietras C, von Palffy S, Askmyr M, Rissler M, Schrappe M, Cario G, Castor A, Pronk CJH, Behrendtz M, Mitelman F, Johansson B, Paulsson K, Andersson AK, Fontes M, Fioretos T. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. Nat Commun 2016;7:11790. (PMID: 27265895)
31. Gaffo E, Boldrin E, Dal Molin A, Bresolin A, Bonizzato A, Trentin L, Frasson C, Debatin KM, Meyer LH, Te Kronnie G, Bortoluzzi S. Circular RNA differential expression in blood cell populations and exploration of circRNA deregulation in pediatric acute lymphoblastic leukemia. Sci Rep 2019;9(1):14670. (PMID: 31605010)
32. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS One 2012;7(2):e30733. (PMID: 22319583)

Tables

Table 1. Outlier expression of isoforms as markers of DNA event

Assay	Gene	Outlier isoform	DNA event	Sensitivity	Specificity
FPH	*IKZF1*	Exon skipping	Δ intragenic or 3'	97.0% (32/33)	97.8% (226/231)
		e3e8	Δ exons 4-7	92.9% (13/14)	98.4% (246/250)
		e1e8	Δ exons 2-7	100% (5/5)	98.8% (256/259)
		e3:g.50505862	Δ exons 4-8	100% (5/5)	100% (259/259)
		e2:g.50505862	Δ exons 3-8	100% (2/2)	100% (262/262)
		e1e5 (no e1e4)	Δ exons 2-4	100% (1/1)	100% (263/263)
		e1e4	Δ exons 2-3	100% (2/2)	100% (262/262)
		e2e4	Δ exon 3	100% (1/1)	100% (263/263)
		e1e3	Δ exon 2	100% (2/2)	100% (262/262)
		e5e5	PTD exon 5 and N159Y	100% (2/2)	100% (262/262)
	ERG	ERGaltA, e4e10, or e4e12	*DUX4*r	90% (9/10)	100% (35/35)
		e4e10 or e4e12	Δ exons 5-9 or 5-11	85.7% (6/7)	100% (257/257)
		e3e8 or e3e5	Δ exons 4-7 or 4	0% (0/3)	100% (261/261)
	PAX5	Exon skipping	Δ intragenic	100% (6/6)	88.5% (23/26)
		e5e2	Gain exons 2-5	100% (3/3)	100% (29/29)
		e7e2	Gain exons 2-7	100% (1/1)	96.8% (30/31)
		e7e7	Gain of exon 7	100% (1/1)	100% (31/31)
	KMT2A	e8e2, e10e2, e8e4, or e6e3	*KMT2A*-PTD	100% (9/9)	100% (255/255)
SBT	*FGFR1*	e18e9* or e18i9*	*FGFR1*-TKDD	100% (7/7)	100% (79/79)

*these breakpoints are intra-exonic or intronic

PTD: partial tandem duplication

Figure Legends

Figure 1. Novel recurrent *IKZF1* isoform junctions. (A) Our annotation-independent split-read approach identified novel junctions connecting *IKZF1* exons 2 or 3 to intergenic breakpoints approximately 25-50 kb downstream of *IKZF1*, thereby compatible with previously characterized DNA breakpoints associated with deletions involving exons 3-8 or 4-8[25,26]. Each of the intergenic breakpoints (chr7:50498139, chr7:50505862, chr7:50525921) was adjacent to AG putative cryptic splice acceptor sites. (B) B-ALL cases N1-N5 and N6-N7 expressed all three exon 3 and both exon 2 isoforms respectively and were associated with deletions of exons 4-8 and 3-8 respectively (Figure 2), thus compatible with alternative splicing where the chr7:50505862 splice site consistently had the dominant usage. All other cases had either no expression or very low level expression of at most one isoform, raising possibility of low-level read-through splicing.

Figure 2. Outlier VAF of *IKZF1* isoforms versus *IKZF1* deletion status. Outlier VAF of eight *IKZF1* exon-skipping isoforms demonstrated an overall sensitivity of 97.0% (32/33) and specificity of 98.2% (226/231) for intragenic or 3' deletion by DNA testing and almost always corresponded to deletion of the skipped exons. The false negative was associated with poor RNA quality; its VAF (0.93%) almost reached the outlier threshold (1.1%) and was higher than every other negative sample (next highest VAF of 0.44%), raising the possibility of a very low-level deletion (possibly from sampling differences) or a deletion with differential RNA degradation between tumor and wild-type cells. Of the 5 false positives, one was an e1e8 outlier associated with a slightly different but incompatible deletion (exons 1-7), thus favoring either incorrectly annotated deletion boundaries or multiple deletions on separate alleles (exon 2-7 and exon 1). The remaining 4 were e3e8 outliers favored to correspond to subclonal deletions below the limit of detection of DNA based methods (Figure 3).

Figure 3. RNA-based VAF of *IKZF1* exon-skipping isoforms versus DNA-based VAF of deletions of the skipped exons. RNA-based VAFs of *IKZF1* e3e8 isoforms were consistently greater than DNA-based VAFs of the underlying deletions of exons 4-8, raising the possibility of RNA allelic overexpression of the mutant allele, however assay bias could not be entirely excluded. Regardless of the source, the behavior enhanced limit of detection by FPH empirically.

Figure 4. *IKZF1*-PTD of exon 5 in B-ALL with N159Y. In 2 of 2 B-ALL cases harboring *IKZF1* N159Y variants, manual review in IGV revealed partial tandem duplications spanning exon 5 with breakpoints in introns 4 and 5 and with lengths (A) 344 bp and (B) 396 bp, corresponding to RNA outlier expression of *IKZF1* e5e5. The breakpoints were sufficiently near exon 5 to be captured fortuitously by RHP baits targeting exon 5 but were undetected by the informatic pipeline due to lack of a structural variant caller and filtering of intronic variants regardless. In both cases, N159Y was present 5' of the mutant junction in some reads and 3' of the mutant junction in other reads, implying duplication of N159Y occurring after N159Y in the clonal hierarchy. The PTDs were described by the following modified notation:

A. c.589+165_589+166insA/422-10_589+165[D1:c.475A>T;D2:c2.475A>T]
B. c.589+160_589+161insTCAC/422-64_589+160[D1:c.475A>T;D2:c2.475A>T]

We also manually reviewed the copy number profiles from RHP, which by convention did not report single exon changes due to noise, however we did not observe gain of exon 5 in either case. The small PTDs may have contributed to the absence of a copy number signal due to loss of supplementary alignment read counts in the CNV algorithm and disrupted UMI reduction in the UMI algorithm.

Figure 5. Outlier VAF of ERGaltA, *ERG* e4e12, or *ERG* e4e10 versus *DUX4* rearrangement status and *ERG* deletion status. Outlier expression of either ERGaltA, *ERG* e4e10, or *ERG* e4e12 was 90% (9/10) sensitive and 100% (35/35) specific for *DUX4* rearrangements as determined by wtRNAseq. Outlier expression of e4e10 or e4e12 was 85.7% (6/7) sensitive and 100% (249/249) specific for intragenic deletions of exons 5-9 or 5-11, where the lone false negative had borderline RNA quality. By contrast, there was zero expression of e3e8 or e3e5 in cases with deletions of exons 4-7 or exon 4.

Figure 6. Outlier expression of *PAX5* out-of-order exon junctions may be sensitive but not specific for PAX5amp. (A) Isoform expression. Four cases (P1-P4) had clear outlier expression of *PAX5* e5e2, e7e2, and/or e7e. Another case minimally exceeded the outlier threshold for *PAX5* e4e2 and harbored a *PAX5* rearrangement involving exon 4, thus favoring altered circRNA expression in the context of a fusion and a suboptimally low threshold. A final case, which barely exceeded the outlier threshold for *PAX5* e7e2, was similarly favored to reflect circRNA and a suboptimal threshold. (B-E) Copy number gains by MLPA. Cases P1 (B) and P3 (D) demonstrated higher order intragenic gains within *PAX5* by MLPA, were not associated with a known molecular subtype, and were favored to represent PAX5amp. P3 demonstrated 4 probably copy number number levels in *PAX5* suggestive of multiple gain events (e1-e7, e2-e7, and e2-e5), thereby consistent with the high outlier e7e2 and moderate outlier e5e2 VAFs. Cases P2 (C) and P4 (E) demonstrated shallow intragenic gains and harbored other rearrangements determining their molecular subtypes

Figure 7. *KMT2A* out-of-order exon junctions as a marker of *KMT2A*-PTD. Outlier expression of *KMT2A* e8e2, e10e2, e8e4, or e6e3 was 100% (9/9) sensitive and 100% (255/255) specific for known *KMT2A*-PTDs of exons 2-8, 2-10, 4-8, or 3-6. The highest VAFs suggested the presence of allelic complexity, such as copy neutral loss of heterozygosity, higher order duplications, or episomal amplifications. Note also that our definition of expressed VAF underestimated the true percentage of *KMT2A*-PTD transcripts since PTD transcripts also contained wild-type junctions causing an overestimated wild-type denominator; the true fraction of *KMT2A*-PTD transcripts was more accurately assessed by the formula 1/((1/VAF)-1). The highest expression levels of e8e2 after *KMT2A*-PTDs occurred in *KMT2A*r cases with fusion breakpoints involving *KMT2A* exon 8, however these never reached outlier status. *KMT2A* wild-type cases occasionally demonstrated a few e8e2 reads, consistent with infrequent low-level circRNA.

Supplementary Figure Legends

SFig 1. Outlier expression of *IKZF1*, *ERG*, or *PAX5* isoforms. In the heme cohort (n=407), outlier expression occurred for 57 isoform junctions across 92 samples consisting of 85 B-ALLs, 1 *BCR-ABL1*-positive MPAL (characterized by a predominant population of B-ALL blasts and a minor population of AML blasts), 4 AMLs (including 2 serial samples from the same patient), 1 remission sample, and 1 sample without a neoplastic diagnosis, thus demonstrating an overall 41.5% (86/207) sensitivity and 96.5% (193/200) specificity for a component of B-ALL. The serial AMLs had outlier expression of *ERG* e6e8 barely exceeding the outlier threshold, suggestive of false positives from alternative splicing and a suboptimally low threshold. Other isoform junctions with lower expression levels were highly correlated with expression of another isoform, consistent with alternative splicing in the context of the same underlying DNA structural event; these included *IKZF1* e1e4 (NM001291846), e1e8 (ENST0000484847:ENST0000331340), e2e5, e2e4 (NM001291847), e3:g.50498139, e3:g.50525921, and e2:g.50498139, ERGaltB and *ERG* e4e11, and *PAX5* e5e3 and e5e4. Comparisons with established and emerging B-ALL molecular subtypes showed various associations including (i) *IKZF1* e5e5 with *IKZF1* N159Y (Fig 4, SFig 3), ERGaltA, *ERG* e4e10, or *ERG* e4e12 with *DUX4*r (Fig 5), and certain *PAX5* deletions with *CRLF2*r or *ETV6-RUNX1* (SFig 5).

SFig 2. *IKZF1* gene expression does not reliably distinguish haploinsufficiency. Cases with broad *IKZF1* shallow (single-copy) deletions were not distinguished from non-deleted B-ALL based on *IKZF1* gene expression under FPH (p=0.36), although a trend appeared for deep *IKZF1* deletions (p=0.06). *IKZF1* gene expression was significantly higher in cases with e3e8 outliers versus non-deleted B-ALL (p=0.0007), raising the possibility of allelic overexpression.

SFig 3. *IKZF1* e5e5 transcripts with duplication of N159Y. Both B-ALL cases with *IKZF1* N159Y (red positions in the reads, all surrounded by black box) demonstrated outlier expression of e5e5 where N159Y was present both 5' of the e5e5 junction (reads surrounded by dark brown boxes) and 3' of the e5e5 junction (reads surrounded by dodger blue boxes). Note that the softclipped ends of the reads (yellow and cyan boxes) match the 5' and 3' ends of exon 5 respectively, consistent with e5e5 junctions.

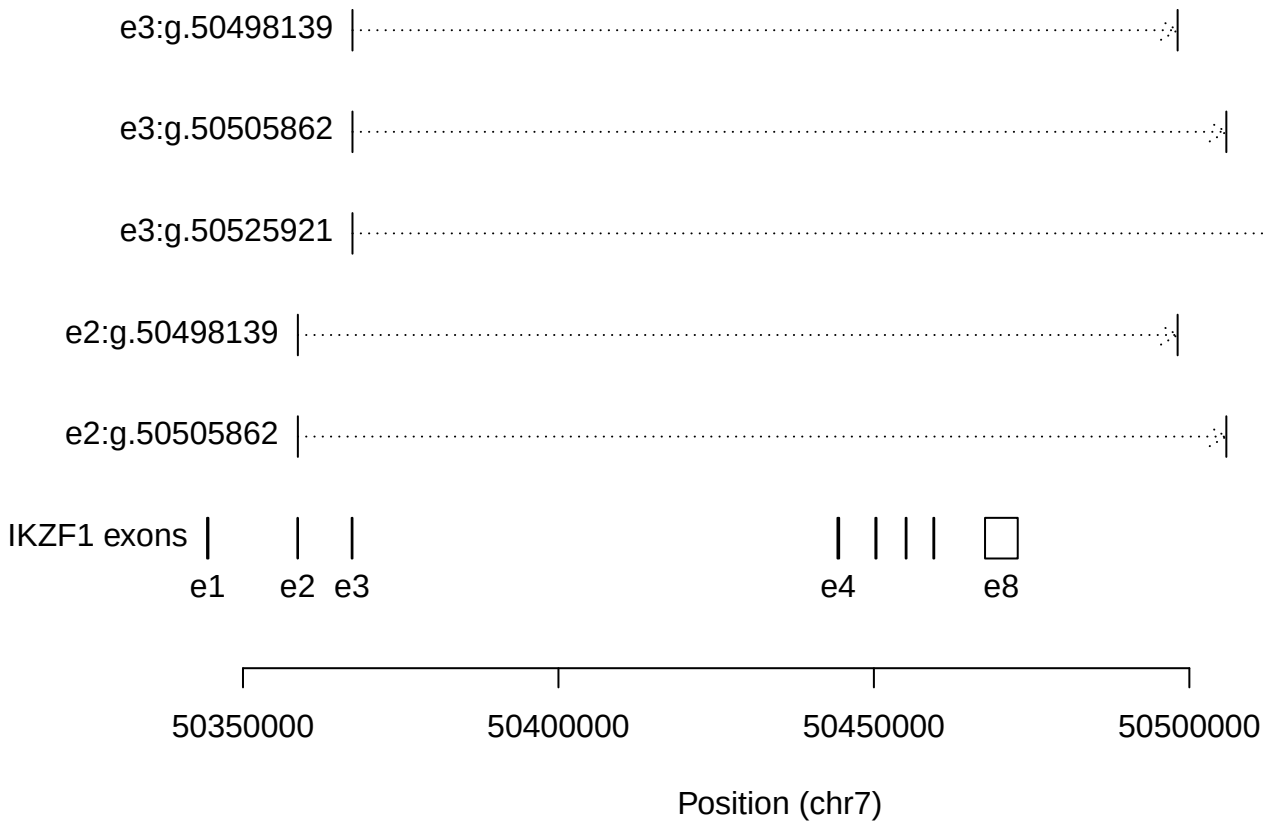
SFig 4. t-SNE of B-ALL cases relative to FPH expression of targeted genes. Under t-SNE applied to FPH gene expression data derived from the limited set of targeted genes, 11/12 B-ALL cases with known *DUX4* rearrangements and/or *ERG* deletions (dodger blue dots surrounded by open black circles) clustered together with 5/6 cases with outlier ERGaltA, e4e10, and/or e4e12 expression but unknown *DUX4*r status (dodger blue dots without surrounding circle). Two cases without outlier expression (black Xs) clustered nearby and did not have a well-defined subtype by molecular or cytogenetic studies; *DUX4*r status was negative by wtRNAseq in 1 of the cases and unknown in the other case. The remaining known *DUX4*r case (separate dodger blue dot surrounded by open circle) had lower cellularity and borderline RNA quality, and clustered nearby other cases with low blast percentage and/or borderline RNA quality. The final case with outlier ERGaltA expression (separate dodger blue dot without

surrounding circle) had the DUX4/ERG signature by LDA card but was notable for its relapse status, suggesting the possibility of altered expression from progression events. Similar overall clustering patterns were observed under hierarchical clustering but were considerably less apparent by PCA or UMAP.

SFig 5. Fusions associated with outlier expression of *PAX5* isoforms. Cases with the highest VAFs (> 50%) of *PAX5* e5e2 or e7e2 were not associated with fusions or cytogenetically defined subtypes, favoring PAX5amp, whereas cases with moderate VAFs (~20%) of e5e2 or e7e7 were associated with a *PDGFRB* fusion and *TCF3-PBX1* respectively. Outlier expression of e5e8, e5e9, e7e9, or e8e10 was associated with *CRLF2* rearrangements while outlier expression of PAX5 e1e6, e1e7, e1e8, or e1e9 was associated with *ETV6-RUNX1*.

SFig 6. *PAX5* exon-skipping isoforms are associated with intragenic deletions. (A) Outlier expression of *PAX5* exon-skipping isoforms. Outlier thresholds were relatively high (close to 40%) for junctions associated with annotated alternative isoforms (e7e9, e8e10), however high VAFs indicative of deletion and LOH were robustly identified. Other thresholds were usually at 1% VAF except for *PAX5* e6e8. Two cases (Q4 and Q5) had outlier expression of 3 isoforms at similar VAFs around 40%-60%. (B-I) Copy number losses by MLPA. Cases Q2 (B) and Q3 (C) had e1e7 VAFs near 50% and corresponding 1-copy deletions of exons 2-6 in ~100% of cells. Cases Q4 (D) and Q5 (E) with multiple isoforms accordingly demonstrated multiple copy number levels consistent with multiple loss events. Cases Q7-Q9 (G-I) had low outlier VAFs (<20%) and were discordantly negative for intragenic deletions by MLPA, however the limit of detection by MLPA has been previously described as 15-30%.

Fig 1A


e3:g.50498139
e3:g.50505862
e3:g.50525921
e2:g.50498139
e2:g.50505862
IKZF1 exons
e1
e2
e3
e4
e8
50350000
50400000
50450000
50500000
Position (chr7)

B

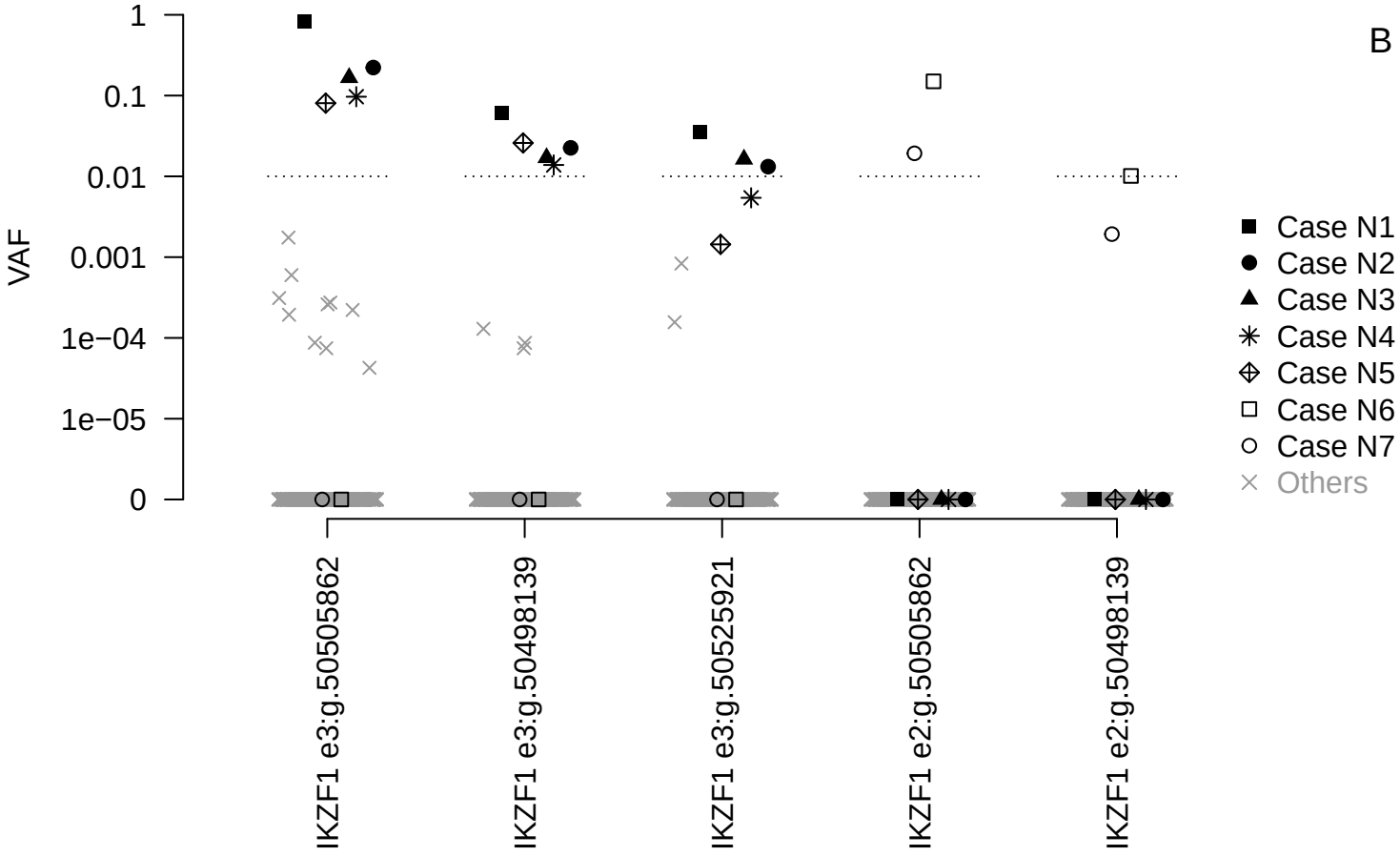
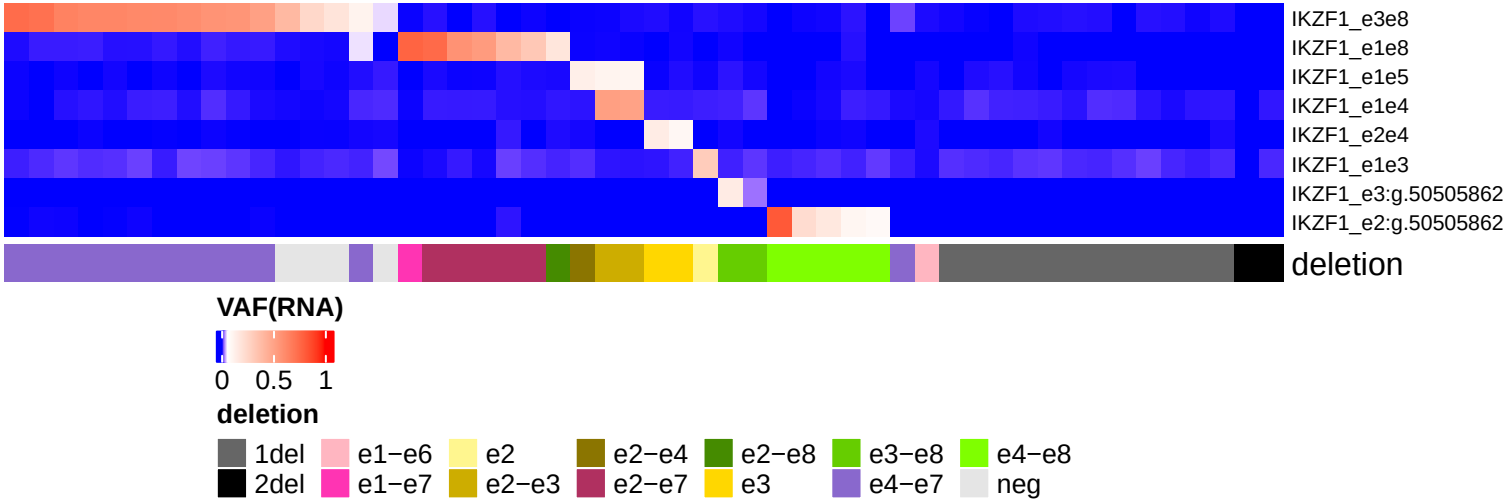

VAF
1
0.1
0.01
0.001
1e−04
1e−05
0
Case N1
Case N2
Case N3
Case N4
Case N5
Case N6
Case N7
Others
IKZF1 e3:g.50505862
IKZF1 e3:g.50498139
IKZF1 e3:g.50525921
IKZF1 e2:g.50505862
IKZF1 e2:g.50498139

Fig 2


IKZF1_e3e8
IKZF1_e1e8
IKZF1_e1e5
IKZF1_e1e4
IKZF1_e2e4
IKZF1_e1e3
IKZF1_e3:g.50505862
IKZF1_e2:g.50505862
deletion
VAF(RNA)
0 0.5 1
deletion
1del
2del
e1−e6
e1−e7
e2
e2−e3
e2−e4
e2−e7
e2−e8
e3
e3−e8
e4−e7
e4−e8
neg

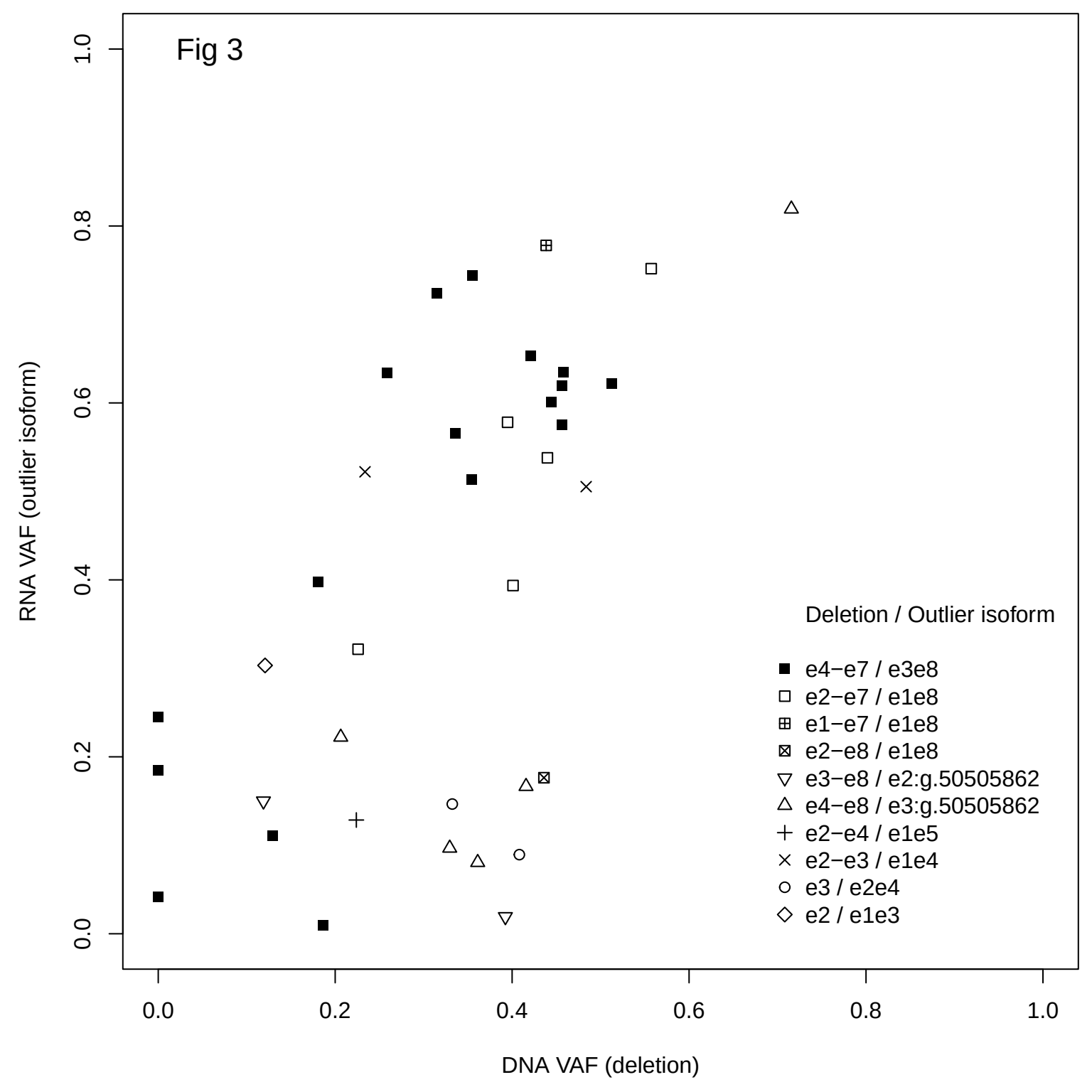
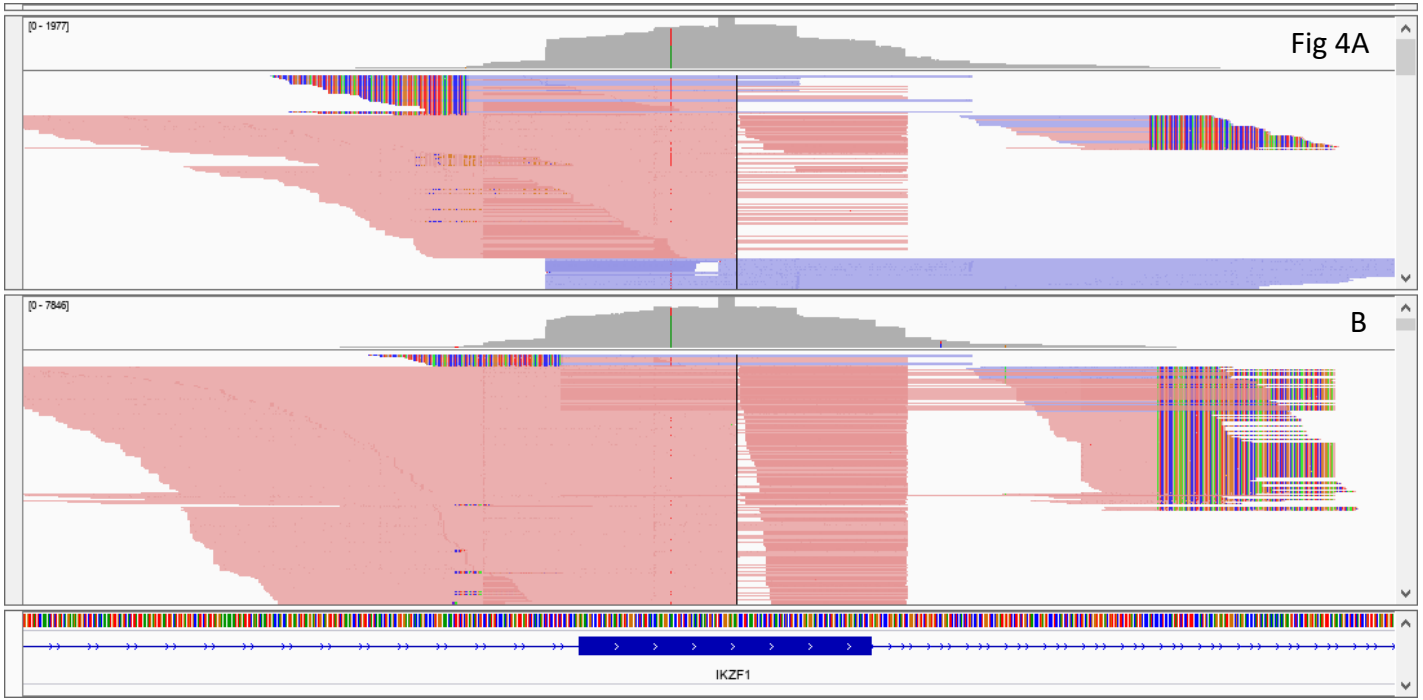


Fig 3
RNA VAF (outlier isoform)
DNA VAF (deletion)
0.0
0.2
0.4
0.6
0.8
1.0
Deletion / Outlier isoform
e4−e7 / e3e8
e2−e7 / e1e8
e1−e7 / e1e8
e2−e8 / e1e8
e3−e8 / e2:g.50505862
e4−e8 / e3:g.50505862
e2−e4 / e1e5
e2−e3 / e1e4
e3 / e2e4
e2 / e1e3


[0 - 1977]
Fig 4A
[0 - 7846]
B
IKZF1

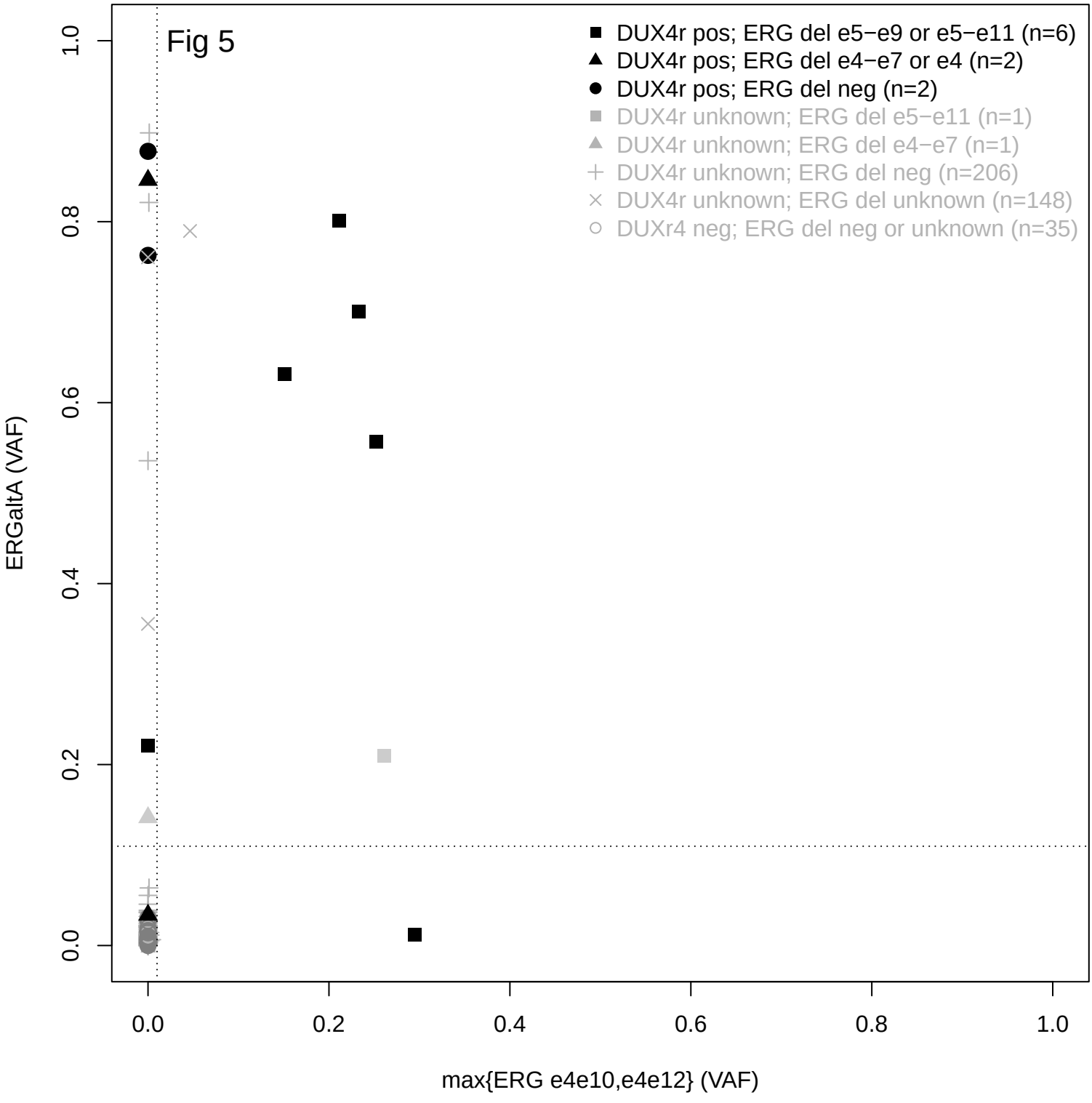


Fig 5
DUX4r pos; ERG del e5−e9 or e5−e11 (n=6)
DUX4r pos; ERG del e4−e7 or e4 (n=2)
DUX4r pos; ERG del neg (n=2)
DUX4r unknown; ERG del e5−e11 (n=1)
DUX4r unknown; ERG del e4−e7 (n=1)
DUX4r unknown; ERG del neg (n=206)
DUX4r unknown; ERG del unknown (n=148)
DUXr4 neg; ERG del neg or unknown (n=35)
ERGaltA (VAF)
0.0
0.2
0.4
0.6
0.8
1.0
max{ERG e4e10,e4e12} (VAF)

Fig 6A
Case P1
Case P2
Case P3
Case P4
Others*
VAF
0.0
0.2
0.4
0.6
0.8
1.0
PAX5_e5e2
PAX5_e7e2
PAX5_e7e7
PAX5_e4e2
B
Case P1
C
Case P2
D
Case P3
E
Case P4
Relative copy number
0.0
0.5
1.0
1.5
2.0
2.5
PAX5 (e10 to e1)
chr9 (other targets)

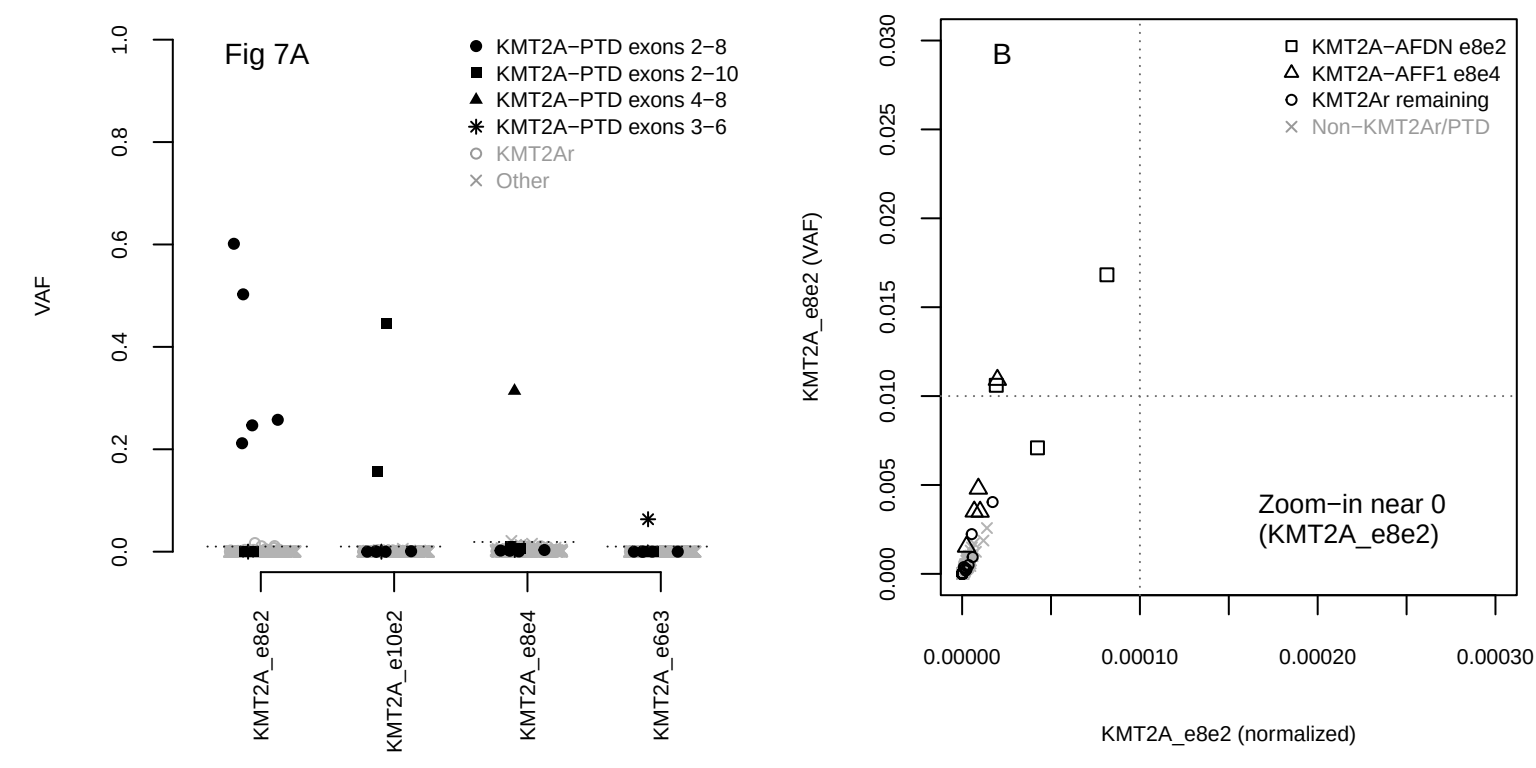
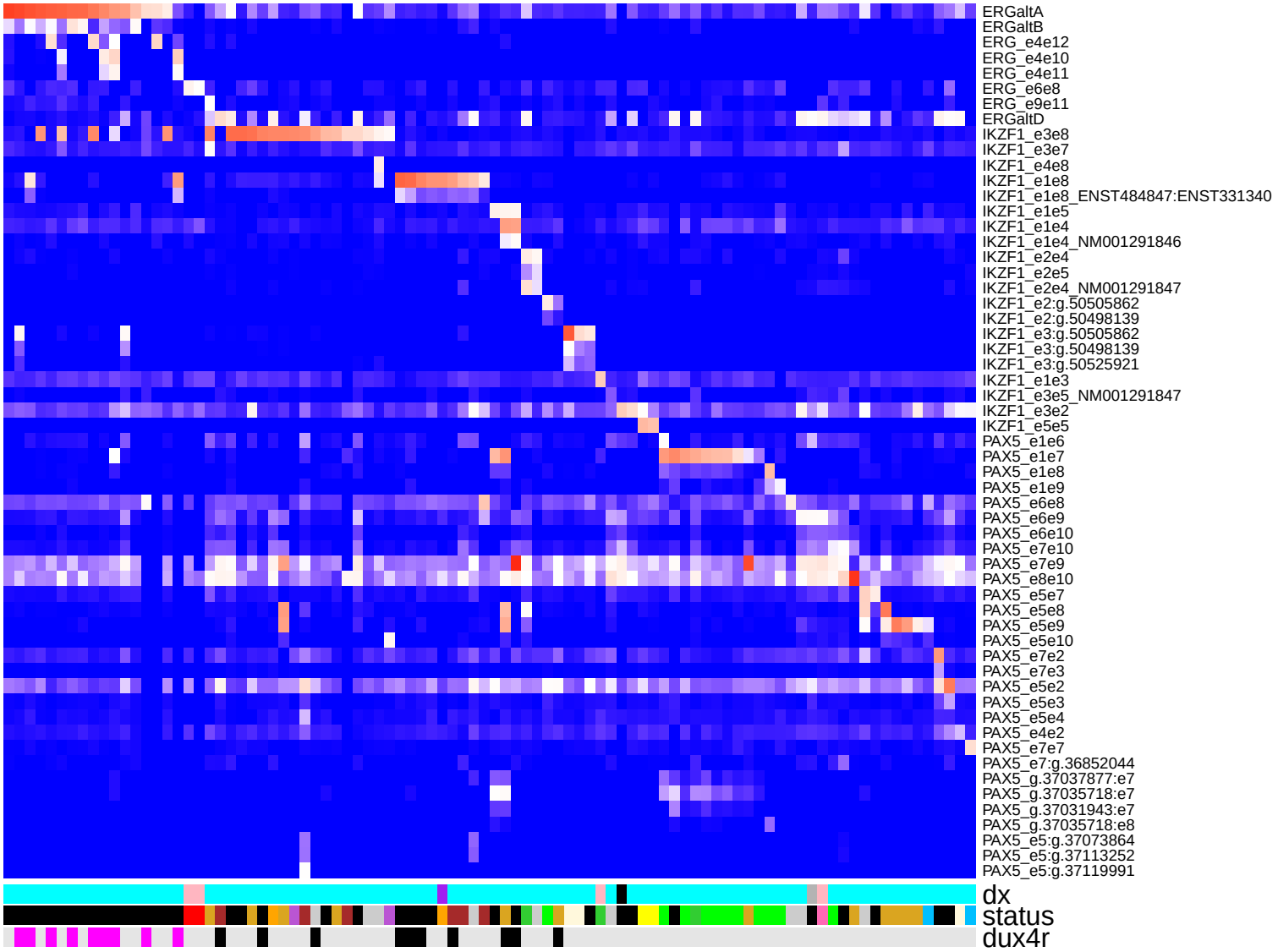


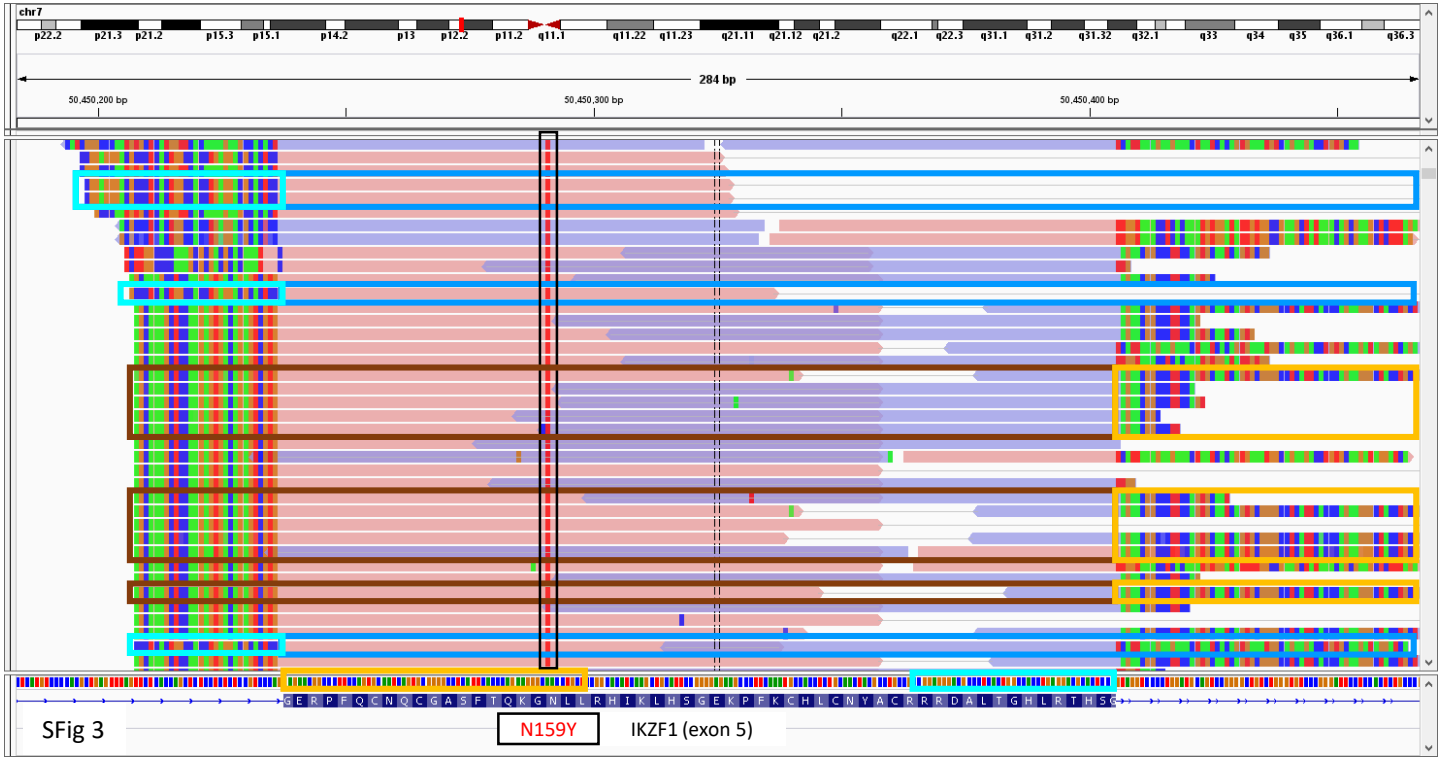
Fig 7A
KMT2A−PTD exons 2−8
KMT2A−PTD exons 2−10
KMT2A−PTD exons 4−8
KMT2A−PTD exons 3−6
KMT2Ar
Other
VAF
0.0
0.2
0.4
0.6
0.8
1.0
KMT2A_e8e2
KMT2A_e10e2
KMT2A_e8e4
KMT2A_e6e3
B
KMT2A−AFDN e8e2
KMT2A−AFF1 e8e4
KMT2Ar remaining
Non−KMT2Ar/PTD
KMT2A_e8e2 (VAF)
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Zoom−in near 0
(KMT2A_e8e2)
0.00000
0.00010
0.00020
0.00030
KMT2A_e8e2 (normalized)

SFig 1


ERGaltA
ERGaltB
ERG_e4e12
ERG_e4e10
ERG_e4e11
ERG_e6e8
ERG_e9e11
ERGaltD
IKZF1_e3e8
IKZF1_e3e7
IKZF1_e4e8
IKZF1_e1e8
IKZF1_e1e8_ENST484847:ENST331340
IKZF1_e1e5
IKZF1_e1e4
IKZF1_e1e4_NM001291846
IKZF1_e2e4
IKZF1_e2e5
IKZF1_e2e4_NM001291847
IKZF1_e2:g.50505862
IKZF1_e2:g.50498139
IKZF1_e3:g.50505862
IKZF1_e3:g.50498139
IKZF1_e3:g.50525921
IKZF1_e1e3
IKZF1_e3e5_NM001291847
IKZF1_e3e2
IKZF1_e5e5
PAX5_e1e6
PAX5_e1e7
PAX5_e1e8
PAX5_e1e9
PAX5_e6e8
PAX5_e6e9
PAX5_e6e10
PAX5_e7e10
PAX5_e7e9
PAX5_e8e10
PAX5_e5e7
PAX5_e5e8
PAX5_e5e9
PAX5_e5e10
PAX5_e7e2
PAX5_e7e3
PAX5_e5e2
PAX5_e5e3
PAX5_e5e4
PAX5_e4e2
PAX5_e7e7
PAX5_e7:g.36852044
PAX5_g.37037877:e7
PAX5_g.37035718:e7
PAX5_g.37031943:e7
PAX5_g.37035718:e8
PAX5_e5:g.37073864
PAX5_e5:g.37113252
PAX5_e5:g.37119991
dx
status
dux4r
pct
0
0.5
1
dx
AML
MPAL
Remission
B−ALL
None
status
BCR−ABL1
CBFB−MYH11
ETV6−RUNX1
HHD
IKZF1_N159Y
None
TCF3−PBX1
CBFA2T3−GLIS2
CRLF2r
ETV6r
iAMP21
Kinase
PAX5r
dux4r
Neg
Pos
Unknown

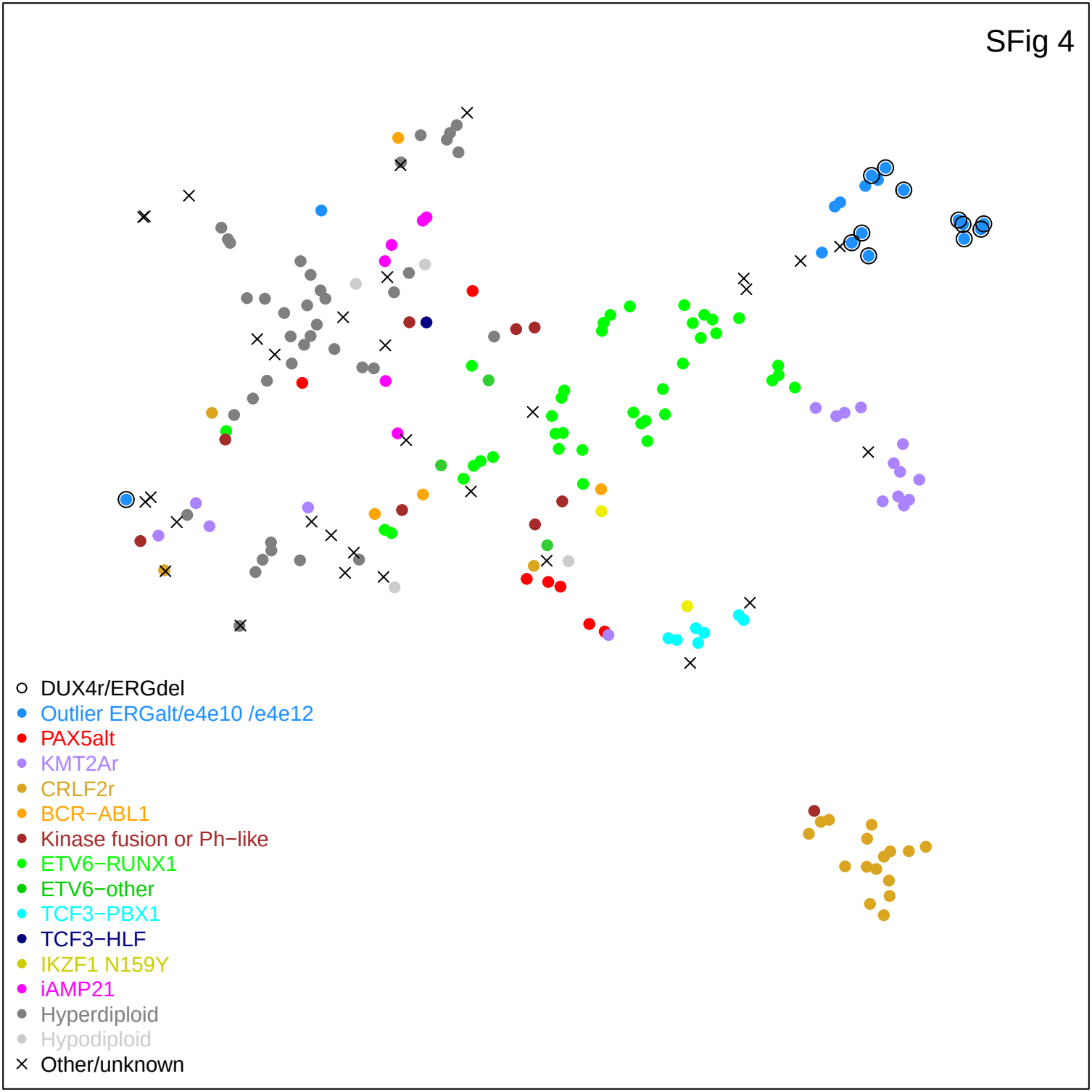
SFig 2

IKZF1 expression (% total)
0
1
2
3
4
5
6
7
e3e8
e1e8
e1e5
e1e4
e2e4
e2:g.50505862
e3:g.50505862
shallow del
deep del
non−outlier + no del
IKZF1 outlier status or deletion status (B−ALL)


chr7
p22.2 p21.3 p21.2 p15.3 p15.1 p14.2 p13 p12.2 p11.2 q11.1 q11.22 q11.23 q21.11 q21.12 q21.2 q22.1 q22.3 q31.1 q31.2 q31.32 q32.1 q33 q34 q35 q36.1 q36.3
284 bp
50,450,200 bp
50,450,300 bp
50,450,400 bp
N159Y
IKZF1 (exon 5)

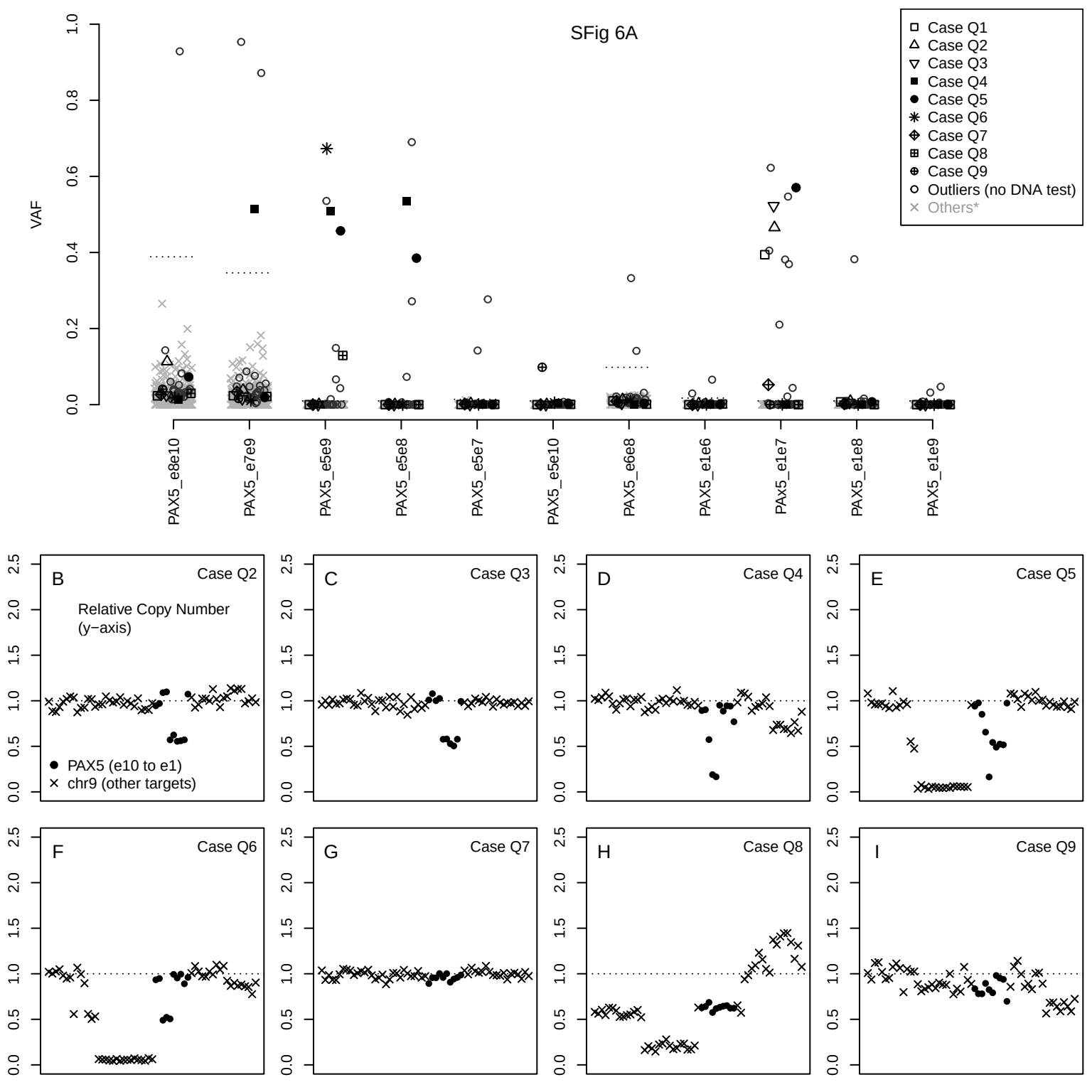
SFig 3

SFig 4


DUX4r/ERGdel
Outlier ERGalt/e4e10 /e4e12
PAX5alt
KMT2Ar
CRLF2r
BCR−ABL1
Kinase fusion or Ph−like
ETV6−RUNX1
ETV6−other
TCF3−PBX1
TCF3−HLF
IKZF1 N159Y
iAMP21
Hyperdiploid
Hypodiploid
Other/unknown

SFig 5


PAX5_e7e2
PAX5_e5e2
PAX5_e4e2
PAX5_e7e7
PAX5_e8e10
PAX5_e7e9
PAX5_e5e9
PAX5_e5e8
PAX5_e5e7
PAX5_e5e10
PAX5_e6e8
PAX5_e1e6
PAX5_e1e7
PAX5_e1e8
PAX5_e1e9
status
pct
0
0.5
1
status
CRLF2r
ETV6−RUNX1
ETV6r
HHD
iAMP21
Kinase
None
PAX5r
TCF3−PBX1


SFig 6A
VAF
Case Q1
Case Q2
Case Q3
Case Q4
Case Q5
Case Q6
Case Q7
Case Q8
Case Q9
Outliers (no DNA test)
Others*
PAX5_e8e10
PAX5_e7e9
PAX5_e5e9
PAX5_e5e8
PAX5_e5e7
PAX5_e5e10
PAX5_e6e8
PAX5_e1e6
PAx5_e1e7
PAX5_e1e8
PAX5_e1e9
B
Case Q2
Relative Copy Number
(y−axis)
PAX5 (e10 to e1)
chr9 (other targets)
C
Case Q3
D
Case Q4
E
Case Q5
F
Case Q6
G
Case Q7
H
Case Q8
I
Case Q9