DEREGULATION OF DUX4 AND ERG IN ACUTE LYMPHOBLASTIC LEUKEMIA

A full list of authors and affiliations appears at the end of the article.

Abstract

Chromosomal rearrangements deregulating hematopoietic transcription factors are common in acute lymphoblastic leukemia (ALL).[1,2] Here, we show that deregulation of the homeobox transcription factor gene *DUX4* and the ETS transcription factor gene *ERG* are hallmarks of a subtype of B-progenitor ALL that comprises up to 7% of B-ALL. *DUX4* rearrangement and overexpression was present in all cases, and was accompanied by transcriptional deregulation of *ERG*, expression of a novel ERG isoform, ERGalt, and frequent *ERG* deletion. ERGalt utilizes a non-canonical first exon whose transcription was initiated by DUX4 binding. ERGalt retains the DNA-binding and transactivating domains of ERG, but inhibits wild-type ERG transcriptional activity and is transforming. These results illustrate a unique paradigm of transcription factor deregulation in leukemia, in which *DUX4* deregulation results in loss-of-function of *ERG*, either by deletion or induction of expression of an isoform that is a dominant negative inhibitor of wild type ERG function.

Address for Correspondence, Jinghui Zhang, Department of Computational Biology, St Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, Jinghui.zhang@stjude.org. Charles G. Mullighan, Department of Pathology, St Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, TEL: 901-595-3387, FAX: 901-495-5947, charles.mullighan@stjude.org.
**Present address: Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA

INTRODUCTION

B-precursor ALL is the commonest childhood tumor, and is a heterogeneous disease comprising multiple subtypes with distinct constellations of somatic structural DNA rearrangements and sequence mutations that commonly perturb lymphoid development, cytokine receptor and Ras signaling, tumor suppression and chromatin modification.[2] However, the genetic basis of a substantial proportion of B-ALL cases remains to be defined. Previous reports identified a subset of B-ALL with a distinct gene expression profile,[3] and frequent deletion of *ERG*, encoding the ETS-family transcription factor v-ets avian erythroblastosis virus E26 oncogene.[4–6] ERG has a key role in hematopoietic differentiation[5,7,8], megakaryopoiesis,[9] and megakaryoblastic leukemia associated with Down syndrome[10–12]. *ERG* is frequently rearranged in carcinoma of the prostate[13] and rarely in acute leukemia[14], and *ERG* overexpression is associated with poor outcome in acute myeloid leukemia[15]. ERG is temporally regulated during B lymphopoiesis (Supplementary Figure 1), suggesting it may regulate B lymphoid development, however its role in the pathogenesis of ALL is unknown.

RESULTS

A subtype of B-ALL with *DUX4* rearrangement and *ERG* deletion

To understand the genetic basis of this subtype of B-ALL, we studied a cohort of 1913 individuals with B-progenitor ALL, including 1347 children, 395 adolescents (age 16–20) and 171 young adults (age 21–39) with B-progenitor ALL using gene expression profiling and analysis of DNA copy number alterations by single nucleotide polymorphism (SNP) arrays in all cases, and whole genome (N=32), exome (N=44) and/or transcriptome sequencing (N=54) in a subset of cases (Supplementary Tables 1–2).

Microarray and transcriptome sequencing data identified 141 (7.6%) ALL cases with a distinct gene expression profile (Figure 1a–b, Supplementary Table 3 and Supplementary Figure 2). This form of leukemia constituted 5.2% of childhood standard risk, 9.4% of childhood high risk, 10.2% of adolescent and 5.4% of adult ALL cases.

Eighty-five (55.6%) of these cases had focal deletions of *ERG* at chromosome 21q22.3 (Figure 1c–d), which were not observed in other B- or T-ALL cases. The *ERG* deletions were confirmed by genomic quantitative PCR and breakpoint mapping, and most commonly involved exons 3–7 (N=27) or 3–9 (N=22) of *ERG* transcript variant 1 (NM_182918.3, the most abundantly expressed transcript in non-B ALL and normal B cells)[5,16]. The presence of conserved heptamer recombinase signal sequences at deletion breakpoints and intervening non-consensus nucleotides indicated that the deletions arise from aberrant recombinase activating gene activity (data not shown). Genomic analysis of a panel of leukemia cell lines showed the B-progenitor cell line NALM-6 exhibited a similar gene expression profile and an intragenic *ERG* deletion (Supplementary Figure 3).

Notably, genes of the double homeobox gene family on chromosomes 4q/10q were among the top upregulated genes (Figure 2a and Supplementary Table 3). Analysis of transcriptome

sequencing data showed that all cases sequenced exhibited rearrangement of *DUX4* to *IGH*, placing *DUX4* under the control of the immunoglobulin heavy chain enhancer resulting in increased expression of *DUX4* (Supplementary Table 2). *DUX4* encodes a double homeobox transcription factor located in a macrosatellite repeat in the subtelomeric repeat unit of chromosome 4q.[17] Deletion of part of this repeat unit is causative of facioscapulohumeral dystrophy, and *DUX4* rearrangements have been reported in a subset of Ewing-like sarcoma (*CIC-DUX4*),[18,19] and recently, a subset of ALL.[20] In each case of this subtype, *DUX4* is inserted adjacent to the *IGH* enhancer region (see Supplementary Note and Supplementary Figure 4a), with variable truncation of the C-terminus of DUX4 and appending of a variable number of amino acids from read-through into the IGH locus (See Supplementary Note, Figure 2b–d). *IGH-DUX4* rearrangement was confirmed in all 6 cases selected for validation by RT-PCR and/or genomic PCR (Supplementary Table 4 and Supplementary Figure 4b–c), and DUX4 overexpression was confirmed by immunoblotting (Figure 2d). Thus, *DUX4* rearrangement is a universal feature of this subtype of B-ALL that exhibits a distinct gene expression profile, with deletion of *ERG* identified in the majority of cases.

The genomic landscape of *DUX4*-rearranged B-ALL

Analysis of genome, exome and transcriptome data demonstrated that in addition to universal rearrangement of *DUX4* and frequent deletion of *ERG*, this subtype of ALL is characterized by a distinct mutational landscape (Figure 3). We identified a mean of 17.5 non-silent sequence mutations per case (range 2–42) and a paucity of structural genetic alterations (Supplementary Tables 5–9). Alterations of lymphoid transcription factor genes were present in 46.5% of cases (*IKZF1* 36.7% and *PAX5* 11.3%). Notably, while *IKZF1* alterations are associated with poor outcome in other subtypes of childhood B-ALL[21,22], they were not associated with poor outcome in childhood *DUX4/ERG* ALL (Supplementary Figure 5). Mutations in transcription factors and transcriptional regulators were observed in 21% of *DUX4/ERG* cases, including *MYC, MYCBP2, MGA* and *ZEB2*, that were uncommon in other subtypes of B-ALL, including 209 representative B-ALL and 16 T-ALL cases subjected to whole genome or exome sequencing (a listing of recurrently mutated and deleted genes across ALL subtypes is provided in Supplementary Table 10). Additional recurrent mutations included those activating Ras signaling (35.2%), cell cycle regulation (22.5%); and epigenetic modifiers (56.3%), most commonly *KMT2D, SETD2, ARID2* and *NCOR1* (Supplementary Figure 6).

ERG deregulation in DUX4-rearranged B-ALL

DUX4-rearranged cases also exhibited profound transcriptional deregulation of *ERG*. Although the absolute level of expression of *ERG* was not higher in comparison to other ALL subtypes, this subtype of ALL was characterized by expression of multiple aberrant coding and non-coding *ERG* isoforms. Using RT-PCR, we detected internally deleted *ERG* transcripts corresponding to the exon 3–7 and 3–9 deletions (Supplementary Figure 7a). Translation of these internally truncated transcripts was predicted to result in a frame shift and premature truncation. However, corresponding N-terminal truncated ERG proteins were not identified on immunoblotting of leukemic cells harboring *ERG* deletions (data not shown). In contrast, immunoblotting using C-terminus specific antibodies identified a 28

kDa ERG protein in 50 (63.2%) of 79 *DUX4/ERG* ALL cases tested, including cases lacking an *ERG* deletion, suggesting an alternate mechanism of *ERG* deregulation (Supplementary Figure 7b).

Of the 54 *DUX4*-rearranged cases subjected to transcriptome sequencing, the majority expressed aberrant *ERG* transcripts initiated in intron 6 (Supplementary Table 11). The most abundantly expressed isoform was initiated from a novel exon whose 3' splice site was located 197 nucleotides proximal to exon 7 (exon 6 alternate, or exon 6 alt; Figure 4a). The genomic boundaries of this non-canonical exon were determined by rapid amplification of cDNA ends (RACE), RT-PCR with primer walking, and analysis of the orientation of RNA-seq read pairs mapping to this region. This confirmed splicing of exon 6 alt to exon 7 and downstream exons, without evidence for splicing of upstream exons into exon 6 alt, indicating that exon 6 alt is a non-canonical first exon (Supplementary Figure 8).

RT-PCR demonstrated in-frame splicing from *ERG* exon 6 alt to exon 7, resulting in expression of an ERG protein with a new N-terminus of 7 amino acids encoded by exon 6 alt followed by exons 7 through 10 of ERG isoform 1 ("ERGalt a"; Figure 4b). The predicted size of this truncated C-terminal ERG protein corresponded to the size of the protein identified by immunoblotting of ERG ALL leukemic cells (Supplementary Figure 7c).

We next systematically analyzed expression of canonical and non-canonical *ERG* transcripts across a range of leukemias and solid tumors using data from the St Jude Children's Research Hospital – Washington University Pediatric Cancer Genome Project. In addition to ERGalt a, we identified a second isoform with coding potential also encoded using *ERG* exon 6 alt, ERGalt b, and two less abundant transcripts that lack coding potential (ERG alt c/d, Figure 4c, Supplementary Table 11). Expression of the coding isoforms, ERG alt a/b, was observed in the majority of ERG ALL cases, and accounted for the majority of transcripts expressed in such cases, but was uncommon and low level in non-ERG cases, such as a minority of *BCR-ABL1* ALL cases (Figure 4d). Analysis of other pediatric ALL (N=290) and non-hematopoietic (N=572) tumors showed that expression of non-canonical *ERG* transcripts were observed in a minority of other B-ALL cases (e.g. Ph+ ALL) but were usually comprised of transcripts without coding potential. In contrast high-level expression of ERG transcripts initiated from this exon was restricted to *DUX4*-rearranged cases, which was confirmed by quantitative RNA-sequencing and quantitative RT-PCR (Supplementary Table 12).

Additional evidence of *ERG* deregulation was observed, with intron retention upstream of *ERG* exon 7 (Supplementary Figure 9). Analysis of total stranded RNA-sequencing data of 12 *DUX4/ERG*, 40 other B-ALL cases and normal B cell progenitors identified expression of a long non-coding RNA proximal to the first exon of *ERG* in the majority of *DUX4/ERG* cases that was restricted to this form of leukemia, and observed in cases with and without *ERG* deletion (Antisense Long non-coding RNA associated with <u>ERG</u>, or ALE; Supplementary Tables 12–13 and Supplementary Figures 10–11).

DUX4 binds and induces deregulation of *ERG*

The finding of *DUX4* rearrangement as a universal feature of this subtype of ALL, and the unique presence of *ERG* deregulation and deletion suggests that the two phenomena are related and may contribute to leukemogenesis. DUX4 is known to induce deregulated expression of many genes, including transcripts utilizing alternate exons.[23]

To examine the relationship of DUX4 expression and *ERG* deregulation, we first analysed previously reported chromatin immunoprecipitation and sequencing (ChIP-seq) data for DUX4 expressed in human myoblasts.[23] This identified a peak of DUX4 binding at the first, non-canonical exon of ERGalt with two DUX4 binding motifs within this 372 nucleotide region (Supplementary Figure 12). Moreover, analysis of RNA-seq data showed that expression of *DUX4* in myoblasts resulted in expression of ERGalt that was not observed in non-*DUX4* expressing cells (Supplementary Figure 13a). Using ChIP-PCR, we observed binding of DUX4 to the first exon of ERGalt in NALM-6 cells (Supplementary Figure 13b). We next performed DUX4 ChIP-seq in NALM-6 and Reh (*ETV6-RUNX1* ALL) cells that confirmed DUX4 binding at *ERG* exon 6 alt (Figure 5a), as well as at additional transcripts deregulated in *DUX4/ERG* ALL (Supplementary Figure 13c). A comprehensive listing of all peaks of DUX4 binding identified by ChIP-seq annotated with *DUX4/ERG* gene expression and ATAC-seq data is provided in Supplementary Table 14.

We systematically analyzed the expression of transcripts utilizing non-canonical first exons in total stranded RNA-seq data, and identified 45 transcripts significantly deregulated in *DUX4/ERG* ALL, of which three (*ERG, NSD1* and *RNGTT*) were bound by DUX4 (Supplementary Table 15). ERGalt was the second most frequently deregulated transcript utilizing a non-canonical first exon, and the most frequently deregulated of those loci bound by DUX4.

To examine the role of DUX4 overexpression in deregulation of *ERG*, we transduced human cord blood hematopoietic cells and the *ETV6-RUNX1* B-ALL cell line Reh with lentiviral vectors expressing *DUX4* alleles corresponding to those identified in patients (DUX4 E415* and Q334*), and empty lentiviral vector as a control. Truncated DUX4 induced expression of ERGalt in both Reh and human CD34 cord blood cells, as shown by both transcript and protein analysis (expression of full length DUX4 was not tolerated in 293T vector producer cells) (Figure 5b–c Supplementary Figure 13d).

Together, these findings suggest that *ERG* deregulation is caused by DUX4 overexpression, which binds to the alternative transcription initiation site in *ERG* intron 6, and that *ERG* deletion is a secondary event occurring in a transcriptionally active locus primed for RAG-mediated deletion. Such deletion of *ERG* may further impair expression of wild-type *ERG*. Consistent with this proposed mechanism and sequence of deregulation, transcription of ERGalt is initiated in *ERG* intron 6, which lies in the common regions of *ERG* deletion (exons 3–7 and 3–9). Thus, in contrast to other key targets of deletion in ALL such as *PAX5* and *IKZF1*, the deleted *ERG* allele cannot encode the putative oncogenic isoform, but rather it is encoded by the non-deleted *ERG* allele.

In addition, some *DUX4/ERG* cases expressing ERGalt lack a clonal *ERG* deletion but have evidence of subclonal deletions. Of the 46 of 54 cases with transcriptome data that express ERGalt, 27 had a clonal *ERG* deletion. However, 7 cases lacking such a deletion had evidence of a subclonal deletion on genomic PCR (Supplementary Table 2), and emergence of clonal *ERG* deletions was observed in xenografted cases, and in cases of relapsed ALL in which the characteristic gene expression profile was present at both diagnosis and relapse, but in which clonal *ERG* deletions were only observed at relapse (data not shown). Together, these data support the notion that *ERG* deletions are secondary genomic events, but if acquired sufficiently early in leukemogenesis may be present as clonal events.

ERGalt inhibits activity of ERG and promotes leukemogenesis

ERGalt lacks the N-terminal pointed domain and central regulatory domains of wild type ERG, but retains the DNA-binding ETS and transactivation domains. Both wild-type ERG and ERGalt exhibited nuclear localization and bound DNA target sequences (Supplementary Figure 14a and data not shown). In a transcriptional reporter assay using an ERG target gene, gpIX[24], ERGalt displayed diminished transactivating activity and acted as a competitive inhibitor of wild type ERG (Supplementary Figure 14b–c).

These results suggest that ERGalt may directly contribute to leukemogenesis, in part by inhibiting the function of wild-type ERG. To explore this, lineage-negative bone marrow from *Arf*[-/−] mice was transduced with retroviral supernatants expressing wild type ERG or ERGalt, and/or an activated NRAS allele and/or IK6, the dominant negative form of IKZF1 commonly observed in human ALL, followed by in vitro culture and transplantation of transduced cells into lethally irradiated recipients. ERGalt and NRAS[G12D] together promoted serial replating of lymphoid cells (Supplementary Figure 15). Consistent with prior reports, expression of wild type ERG induced a lethal erythromegakaryoblastic leukemia (Figure 6a–d)[25]. In contrast, mice transplanted with ERGalt expressing cells developed lymphoid precursor, biphenotypic or pre-B cell leukemias with longer latency (Figure 6a–d), indicating ERGalt directly promotes lymphoid leukemogenesis.

Discussion

These data provide comprehensive genomic characterization of a subtype of B-progenitor ALL characterized by a distinct gene expression profile and deregulation of two transcription factors, *DUX4* and *ERG*. Our findings indicate that *DUX4* rearrangement is present in all cases with this distinct gene expression profile, and is a clonal event acquired early in leukemogenesis. *DUX4* is not expressed during normal mouse or human B cell development, and translocation to *IGH* provides a mechanism for hijacking into the B cell lineage, as we have previously described for other genes rearranged to antigen receptor loci such as *CRLF2* and *EPOR*.[26–29]

The striking deregulation and deletion of *ERG* is unique to this subtype of ALL. While multiple prior studies have reported *ERG* deletions in B-ALL,[22,30–32] including in *DUX4*-rearranged cases,[33] several of these studies have identified only cases harboring the common deletions and used these as a surrogate to identify a subset of ALL cases with favorable prognosis.[30,31] In contrast, we show that *ERG* transcriptional deregulation is a hallmark of

this subtype, with diverse *ERG* deletions in a majority, but not all cases. Multiple mechanisms of *ERG* genomic alteration were observed. These included expression of aberrant *ERG* transcripts in all cases, including non-canonical coding transcripts, deregulation with intron retention, and long non-coding RNA expression, with expression of the novel coding ERG transcript, ERGalt, in the majority of cases; and clonal or subclonal *ERG* deletions also observed in the majority of cases. We have shown that DUX4 directly binds to the *ERG* locus at the first, non-canonical exon of ERGalt, and in multiple hematopoietic and non-hematopoietic cell types directly induces expression of ERGalt. The deregulation of ERG by DUX4 is reminiscent of deregulation of other ETS family genes in solid tumors,[19] and represents a new mechanism of ETS gene deregulation in leukemia. Moreover, as ERG may represent one of potentially many genes deregulated by DUX4, we performed a systematic, genome-wide, integrated analysis of gene expression, expression of genes utilizing non-canonical first exons, DUX4 ChIP-seq and ATAC sequencing and show that ERGalt is the top deregulated gene in this form of leukemia that is bound by DUX4 and utilizes a non-canonical first exon.

The observations that *ERG* is deleted in only a subset of cases, and that expression of ERGalt is directly induced by *DUX4* suggests that *ERG* deregulation is an important, but secondary event in leukemogenesis. In addition, we have shown that for the majority of cases harboring an *ERG* deletion, deregulation of the *ERG* locus involves both alleles, and cannot be explained by perturbation of a single copy of the gene. The common *ERG* deletions remove intron 6 which harbors *ERG* exon 6 alt, the first exon encoding ERGalt, and the region bound by DUX4. Thus, in a case with a clonal *ERG* deletion, ERGalt must be expressed from the non-deleted *ERG* allele. Thus, our data suggest a sequence of events in which rearrangement of *DUX4* is an early, leukemia-initiating event that results in binding to the *ERG* locus and deregulated expression of coding and non-coding transcripts. The resulting increased chromatin accessibility evidenced by ATAC-sequencing renders the locus susceptible to RAG-mediated deletion of *ERG*. Should this occur early in leukemogenesis, *ERG* deletion is observed as a clonal event, but if later, a subclonal event that may become clonal during disease progression (e.g. at relapse, or as a surrogate, passage in an immunocompromised mouse).

The interplay of *DUX4* and *ERG* deregulation in leukemogenesis will require detailed future examination in appropriately engineered mouse models that must account for the fact that ERG exon 6 alt is incompletely conserved in the mouse, and primary mouse hematopoietic cells were thus not suited to examining induction of ERG alt by DUX4. That notwithstanding, the findings of inhibition of the transcriptional activity of wild-type ERG by ERGalt, and documentation of aberrant intron retention and/or deletion of *ERG* in all cases in this form of leukemia, but rarely in any other tumor indicate that inhibition or loss of ERG activity is required in the pathogenesis of human *DUX4/ERG* ALL. The notion that ERGalt sustains lymphoid colony replating, and induces leukemia with longer latency than wild type ERG, but with a shift to a biphenotypic or lymphoid lineage, further supports this notion.

These findings have important clinical implications, as *DUX4/ERG* ALL is associated with favorable outcome, irrespective of the presence of concomitant genetic alterations otherwise

associated with poor outcome in other contexts, such as deletion of *IKZF1*. *DUX4* rearrangement is not evident on karyotypic analysis, and is challenging to identify on analysis of genome and RNA-sequencing data due to the repetitive nature of both breakpoints and the nature of *DUX4* insertion into the *IGH* locus. Thus, in contrast to prior studies that have used identification of only the common (and clonal) *ERG* deletions as a surrogate for identification of a distinct type of leukemia,[30,31] future studies must move beyond identification of *ERG* deletions alone to identify this form of leukemia. All cases harbor rearrangement of *DUX4*, detection of which requires transcriptome and/or genome sequencing. Quantitation of *DUX4* may also identify cases with rearrangements, but must be performed carefully in view of the highly paralogous nature of the *DUX* family of genes. Conventional cytogenetic approaches such as fluorescence *in situ* hybridization are challenging due to the repetitive nature of *DUX4* locus. Identification of such cases, however, is important and to accurately assign risk and guide therapy.

ONLINE METHODS

Patients and samples

Diagnosis and remission samples were obtained from St Jude Children's Research Hospital, the Children's Oncology Group, the Alliance – Cancer and Leukemia Group B, the Eastern Cooperative Oncology Group and the MD Anderson Cancer Center. We examined leukemia samples obtained at diagnosis from 1913 children (age up to age 16), adolescents (age 16–21) and young adults (age 21–39) with B-progenitor ALL (Supplementary Table 1) using microarray gene expression profiling and single nucleotide polymorphism array analysis. The study was approved by the St Jude Institutional Review Board and informed consent was obtained from patients or legal guardians.

Genomic analysis

Whole genome, whole exome, and transcriptome sequencing, and gene expression and SNP array analysis were performed as previously described.[29,30] Additional details of bioinformatic analysis are provided in the Supplementary Note. The genomic landscape was summarized using ProteinPaint[34].

Whole genome and/or exome sequencing was performed for *ERG*-altered B-ALL cases (N=72 cases), *ETV6-RUNX1* ALL (54), ALL with high hyperdiploidy (39), *BCR-ABL1* ALL (39), ALL with low hypodiploidy or near haploid karyotypes (N=19), *TCF3-PBX1* ALL (N=17), *BCR-ABL1*-like ALL (17), miscellaneous B-ALL (17), T-lineage ALL (N=17) and B-ALL with rearrangement of *CRLF2* (7). Paired end whole genome sequencing of tumor and normal DNA was performed using HiSeq 2000 genome sequencers (Illumina) as previously described to at least 30 fold haploid coverage. Exome sequencing was performed using Truseq exome capture baits (Illumina) and GAIIx or HiSeq 2000 sequencers as previously described[35]. Sequence mapping and variant calling was performed as described[35,36]. Data were visualized in ProteinPaint.[37]

Transcriptome sequencing was performed for 175 B-ALL cases including *ERG*-altered cases (54 cases), *ETV6-RUNX1* (N=54), *BCR-ABL1*-positive and *BCR-ABL1*-like ALL (N=27

each), hypodiploid ALL (N=8), high hyperdiploid and miscellaneous ALL (N=5). Total RNA was extracted from leukemia cells using TRIzol (Life Technologies, NY). Total RNA quality and quantity were assessed on Agilent RNA6000 Chip (Agilent, CA) and Qubit (Life). RNA-Seq was prepared from 1 μg total RNA following Illumina RNA-Seq protocols including DNase treatment and Phenol purification, PolyA+ RNA selection by using Oligo-dT beads, cDNA conversion, fragmentation by Covaris Ultrasonicator (Covaris, MA), end repairing, deoxyadenosine tailing, adaptor ligation and PCR amplification (10 cycles). The library with 10 pM was clustered on Illumina cBot and the flowcell was loaded on HiSeq for sequencing using Illumina 2×100bp sequencing kit (Illumina, CA).

Transcriptome sequencing data were mapped to the human genome (hg19) using in-house software, and the resulting alignments were analyzed to identify evidence of both known and novel splicing events. Predicted splice junction sites were post-processed with a software package called RNApeg to correct mapping ambiguities and apply minimum quality control requirements to novel junction calls. RNApeg evaluates junctions for microhomologous mapping ambiguity versus reference junctions in the refGene, ENSEMBL, AceView, and UCSC known gene databases, also correcting against novel exon skips and single-edge matches in those isoforms. Novel junctions without any available reference anchoring are compared with each other and across samples, facilitating standardized comparison. Where ambiguity was identified, coordinates are adjusted and supporting read evidence is combined, producing a more compact and consistent set of junction calls.

The transcript expression levels in transcriptome sequencing data were estimated as Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Briefly, the read counts in the GENCODE annotated gene model were obtained with the HTseq-count program.[38] The FPKM values were computed by normalizing the obtained read counts of the transcripts or genes with the length of the transcripts or genes and the total mapped reads.

A gene was called "expressed" in a given sample if it had a FPKM value 0.35 based on the distribution of FPKM gene expression levels and filtered out genes that were not expressed in any sample from the final gene expression data matrix for downstream analysis. Similar to expression arrays, *limma*[39] with estimation of false-discovery rate (FDR)[40] was also performed between ERG and other B-ALL samples (FDR<0.05).

Total stranded RNA sequencing was performed for 6 B lymphoid progenitor samples flow sorted from human bone marrow, 12 ERG ALL and 40 non ERG B-ALL cases. Expression levels of ERGalt and ALE were examined in 922 hematopoietic, brain and solid tumor cases sequenced by the St Jude Washington University Pediatric Cancer Genome Project, and adult tumor cases sequenced by The Cancer Genome Atlas (non-small cell adenocarcinoma, N=308, squamous non-small cell lung cancer, N=279, low grade glioma, N=467, glioblastoma multiforme, GBM N=167 and ovarian cancer, N=15).

ATAC sequencing

ATAC (assay for transposase-accessible chromatin) sequencing was performed using cell lines (NALM-6, Reh) and xenografts of *DUX4/ERG* ALL as previously described with

minor modifications [41]. NALM-6 and Reh were obtained from the Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures and were subjected to SNP array and transcriptome sequencing, and tested for *Mycoplasma spp.* contamination prior to use. Briefly, 50,000 cells were lysed and isolated nuclei were resuspended in 2× TD buffer. Transposase reactions were purified using MinElute columns (Qiagen) and library fragments amplified using NEB Next HiFi 2× MM (New England Biosciences), with quantitative PCR cycle optimization performed using Kapa SybrFast 2× MM (Kapa Biosystems). Post amplification products were purified using two rounds of incubation with the Agencourt Ampure XP SPRI beads at a 1.4× ratio (Beckman Coulter). The libraries were quantified via Qubit HS DNA kit (Life Technologies), and evaluated for library size distribution on a Bioanalyzer 2100 (Agilent Technologies). Sequencing was performed on a HiSeq 2000 (Illumina) generating paired 100 nucleotide reads.

Reads were aligned to hg19 with BWA with default parameters after trimming illumina adapter sequences. Observed fragment size showed similar enrichment of nucleosome free, mononucleosome, dinucleosome, trinucleosome as described.[41] Reads were adjusted by transposon inserts offset, and the unique aligned nucleosome free fragments of less than 100 nucleotides in size were extraacted. Wig files were generated by extending these nucleosome free fragments to 80bp from the center and uploaded to UCSC Genome Browser for visualization. For annotation of DUX4 binding sites, we called nucleosome free regions considering every nucleosome free fragments as two single-end reads and called nucleosome free regions with MACS2[42] (version 2.1.0.20150603; default parameters with "--extsize 200 --nomodel").

Establishment of *DUX4/ERG* xenografts

Eight xenografts of human *DUX4/ERG* ALL were established using NOD.Cg-*Prkdc*scid *Il2rg*tm1Wjl/*SzJ* (NOD-SCID gamma-null, or NSG) mice.[43] Primary leukemia cells were intravenously injected in NSG mice and engraftment monitored by flow cytometric analysis of peripheral blood using antibodies against human CD19 and CD45. Engraftment was monitored by retro orbital bleeding after 4 weeks, and then every 2–4 weeks as necessary and xenograft cells were harvested when levels of peripheral blood engraftment exceeded 50%.

Cloning and expression of DUX4 in human cord blood samples

All cord blood samples were obtained with informed consent according to procedures approved by the institutional review boards of the University Health Network, Trillium Health Centre, Brampton Civic Hospital, and Credit Valley Hospital, Ontario, Canada. Mononuclear cells were obtained by centrifugation on Lymphoprep medium (Stem Cell Technologies) and were depleted of Lin+ cells (lineage depletion) by negative selection with the StemSep Human Progenitor Cell Enrichment Kit according to the manufacturer's protocol (Stem Cell Technologies). Lin− CB cells were stored at −150°C. Lin− cells were thawed by drop-wise addition of X-VIVO 10 (Lonza), 50% HyClone Cosmic Calf Serum, 100 μg/ml DNase (Roche) and were cultured at a cell density of $1.6 - 2.5 \times 10^6$/mL in X-VIVO 10 medium supplemented with 1% BSA (Roche), L-glutamine (GIBCO), Pen/Strep

(GIBCO), and the following cytokines (all from Miltenyi): SCF (100 ng/ml), Flt3L (100 ng/ml), TPO (50 ng/ml) and IL7 (IL-7; 10 ng/ml).

For lentiviral overexpression studies a pRRL based and Gateway® (ThermoFisher) adapted lentiviral vector was used in which transgene expression is driven by the SFFV promoter and tagBFP by a chimeric EF1α/SV40 promoter (J.E.D., K.K., ST., E.W., manuscript in preparation). The cDNA of wild type and truncated *DUX4* were PCR amplified from leukemic cell RNA and were inserted via Gateway® LR clonase reaction according to manufacturer's instructions (ThermoFisher). VSV-G pseudotyped lentiviral vector particles were produced by polyethyleneimine (PEI) based co-transfection of 10.5 μg pMD.G2, 20.5 μg pCMVR8.74 (both Addgene) and 38 μg transfer vector into HEK-293T cells[44] and titrated on MOLM-13 cells. For transduction, pre-stimulated (15 hrs) CB Lin- cells $(4 \times 10^5/\text{mL})$ were exposed to virus for 32 hrs at an MOI of 4.5 resulting in 22–78% BFP+ cells at day 4 post transduction when cells were sorted on FACSAria II (Becton Dickinson) to obtain BFP+ and BFP– fractions for RNA (5×10^4) and protein lysates $(0.9\text{-}5.6 \times 10^6)$.

Chromatin immunoprecipitation assays

ChIP assays were carried out as described previously.[45] Chromatin shearing was performed according to the truChIP shearing protocol (Covaris; http://covarisinc.com/wp-content/uploads/pn_010179.pdf) with minor modification. Briefly, 25×10^6 Reh and NALM6 cells were incubated for 10 min in 1% formaldehyde in phosphate buffered saline at room temperature, quenched by adding 1/10 volume of 2.5 M glycine. The cells were then washed three times by cold phosphate buffered saline containing proteinase inhibitors and lysed on ice for 10 minutes in lysis buffer (50 mM HEPES PH7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton-X100). Chromatin was washed twice in washing buffer (10 mM TrisCl, pH 8, 200 mM NaCl,1 mM EDTA, 0.5 mM EGTA), then twice in shearing buffer (0.1% SDS, 10 mM TrisCl, pH8, 1 mM EDTA) before resuspension in 1 mL shearing buffer. Chromatin was sonicated in 1 ml AFA millitubes using a Covaris E210 instrument for 20 mintues at 5% duty cycle, intensity 4, 200 cycles per burst at 7°C. The sheared chromatin was spun down 10 minutes at 13,000 rpm, 4°C and the supernatant was mixed with equal amount of ChIP-dilution buffer (0.1% SDS, 30 mM TrisCl, pH 8, 1 mM EDTA, 300 mM NaCl) before ChIP experiments. Immunoprecipitation was performed with a DUX4 antibody (ab124699) and a normal rabbit IgG control (Santa Cruz).

Real-time PCR (Ct method) was employed to determine the enrichment of human *ERG* locus by antibodies for DUX4. The signals were normalized to input as percentile. The primers used are shown in Supplementary Table 16.

To prepare ChIP-seq libraries, 10 ng of ChIP DNA was end-repaired and adapter ligation performed using the NEBNext® ChIP-Seq Library Prep Reagent Set for Illumina® (New England Biosciences). Libraries were purified after 14 rounds of PCR amplification with Q5 DNA Hot Start polymerase (New England Biosciences). Each ChIP-seq library underwent 50-cycle single-end sequencing using TruSeq SBS kit V3 on the Illumina Hiseq 2000.

Sequence reads were aligned to human genome hg19 (GRCh37) by BWA[46] (version 0.5.9-r26-dev using default parameter), and duplicated reads were then marked with Picard

(version 1.65(1160)) and only non-duplicated reads retained for analysis (using SAMtools[47] (parameter "-q 1 -F 1024" version 0.1.18 (r982:295)). Cross-correlation plots were generated by a non-duplicated version of SPP[48] (version 1.11) for quality control (QC) and fragment sizes estimated using caTools (version 1.17) and bitops (version 1.0–6) implemented in R (version 2.14.0). The best fragment size estimated by the cross-correlation plot were used to extend each read to generate bigwig files for viewing in Integrative Genomics Viewer[49] (version 2.3.40), with scaling of the height of bigwig tracks by normalization to 15 million non-duplicated reads. All results of quality control analysis and visualization indicated our data passed ENCODE criteria except for Reh DUX4-ChIP seq, where results were similar to input DNA used as a conrol (relative strand correlation < 1) which was as expected due to the lack of DUX4 expression in this cell line. For annotation, MACS2 was used for peak calling (again using input DNA as a control) and only peaks overlapping in the two replicates for NALM6 were retained and merged as finalized peaks. MAST[50] in the MEME suite[51] (version 4.10.2) was used for scanning DUX4 motifs (JASPAR[52] accession MA0468.1) within finalized NALM6 DUX4 ChIP-seq peaks.

Fluorescent in situ hybridization

Fluorescence in situ hybridization for disruption of the *ERG* locus was performed for 10 novel BALL cases in the training cohort using diagnostic bone marrow or peripheral blood leukemic cells in Carnoy's fixative as previously described[32]. BAC clones RP11-50G3 (5' of *ERG*) and RP11-720N21 (3' of *ERG*) were labelled with rhodamine and fluorescein isothiocyanate, respectively. At least 100 interphase nuclei were scored per case.

RT-PCR and cloning

Wild type ERG, ERGalt and ALE were amplified using Advantage 2 DNA polymerase (Clontech, Mountain View, CA) as previously described[32]. PCR primers are listed in (Supplementary Table 16). PCR products were purified, and sequenced directly and after cloning into pGEM-T-Easy (Promega) and subcloned into MSCV-IRES-GFP retroviral vectors.

Luciferase assays

Twenty four hours after plating 20,000 293T cells per well in a 96-well plate, cells were transfected with equimolar amounts of ERG wild type (MIG-ERG isotype 1), mutant (MIG-ERG-e6alt) or MIG empty vector along with 250 ng of pGL3-gpIX luciferase reporter plasmid and 50 ng of pRL-TK *Renilla* luciferase plasmid DNA (Promega) using Fugene 6 Transfection Reagent (Roche Diagnostics, Alameda, CA). Competition assays were performed by transfecting increasing amounts of mutant ERG plasmid together with a fixed amount of wild type ERG plasmid or vice versa complemented with empty vector. Forty-eight hours post-transfection, cell lysis and measurement of firefly and *Renilla* luciferase activity was performed using the Dual-Glo® Luciferase Assay System (Promega) according to the manufacturer's instructions. Transfections were performed in triplicate in at least two independent experiments. The firefly luciferase - *Renilla* luciferase ratio was reported as mean ± s.e.m.

Immunofluorescence and immunoblotting

Cytospins of pre-B cells were fixed with 4% paraformaldehyde and washed with phosphate-buffered saline (PBS), followed by a 30 minute incubation in a blocking/permeabilization solution of 10% normal goat serum (NGS)/0.1% Triton-X 100/PBS, and then incubated for one hour with primary ERG antibody (Abcam EPR3864) diluted in 3% NGS/0.1% Triton-X 100/PBS. Slides were washed three times in PBS, and then incubated for 45 minutes in 3% NGS/0.1% Triton-X 100/PBS containing a secondary antibody conjugated to Ig-Alexa Fluor 555 (Invitrogen). Slides were washed three times in PBS and mounted with Vectashield (Vector Labs) containing 4'-6-diamidino-2-phenylindol (DAPI). All steps were carried out at room temperature. Images were captured using a Nikon C2 confocal fluorescence microscope.

Immunoblotting of cell line and leukemic cell whole cell lysates was performed as previously described[32,53] using ERG N-terminus specific antibodies (C20 and H95, Santa Cruz Biotechnologies, Santa Cruz, CA), and a C-terminus specific rabbit monoclonal antibody (Abcam 133264). DUX4 immunoblotting was performed with a C-terminus specific (E5-5, ab124699, Abcam) and N-terminus specific (aa82–131, LS-C205474, Lifespan Biosciences, Inc.) antibodies.

Gene Transduction and Transplantation of Lineage-negative Bone Marrow Cells

Mice were housed in an American Association of Laboratory Animal Care (AALAC)-accredited facility and were treated on Institutional Animal Care and Use Committee (IACUC)-approved protocols in accordance with NIH guidelines. For *in vivo* lineage-negative cell transplantation experiments, bone marrow from 8-week-old *Arf*$^{-/-}$ mice[54] was extracted from tibiae and femora. Red blood cells were lysed and the remaining bone marrow cells were incubated with a cocktail of biotinylated anti-mouse antibodies (Gr-1, B220, Ly-1, Ter119, Mac-1; BD Biosciences, San Jose, CA) followed by mixing with streptavidin-coated beads (Dynabeads M-280 Streptavidin, Dynal 112.16; Life Technologies, Grand Island, NY). Cells were separated on a magnet, and unbound cells were collected and incubated at 37°C, 5% CO_2 for two days in the presence of IL-3, IL-6, SCF, IL-7, and Flt3 cytokines (Peprotech; Rocky Hill, NJ). Cells were retrovirally-transduced with MIG-ERG WT or MIG-ERGalt on RetroNectin (Takara Bio, Otsu Shiga, Japan) for 48 hours priors to sorting for GFP-positive cells (BD FACSAria, BD Biosciences, San Jose, CA). Recipient 10-week-old wildtype C57Bl/6 mice were lethally irradiated (11 Gy) 24 hours prior to transplantation with 200,000 GFP-positive cells via tail-vein injection. Animals were not randomized prior to transplantation. Animals were monitored, without blinding, for the development of leukemia clinically by measurement of GFP expression in peripheral blood and sacrificed when moribund, or at the end of the study. Post-mortem flow analysis for B220, CD19, Mac-1, Gr-1, Ter119, and Thy1 was performed on the GFP-positive population of bone marrow and spleen samples to determine lineage of disease. GraphPad Prism v.6 software was utilized to generate Kaplan-Meier survival curves and Mantel-Cox p values for pairwise comparisons of cohorts.

Lineage-negative enrichment and colony assays of murine haematopoietic cells

Bone marrow mononuclear cells were harvested from 10-week old $Arf^{-/-}$ mice and labeled with biotin-conjugated lineage antibodies (Ly-6G and Ly-6C, CD11b, CD45R/B220, CD5, TER-119; BD Biosciences, San Jose, CA), followed by incubation with streptavidin coated magnetic beads (Dynabeads M-280 Streptavidin; Thermo Fisher Scientific Inc., Franklin, MA). Lineage-negative cells were purified by magnetic separation and cultured for 48 hours in Iscove's Modified Dulbecco's Medium/20% fetal bovine serum supplemented with penicillin-streptomycin, L-glutamine, recombinant mouse stem cell factor (SCF; 50 ng/ml), FLT-3 ligand (40 ng/ml), IL-6 (30 ng/ml), IL-3 (20 ng/ml), and IL-7 (10 ng/ml) (Peprotech, Rocky Hill, NJ). Cells were infected on RetroNectin-coated plates for 48 hours (Takara Bio Inc., Shiga, Japan) with MSCV-IRES-GFP (MIG) and MSCV-IRES-RFP (MIR) retroviruses expressing the following: wild-type, $NRAS^{G12D}$, ERG^{wt}, ERG^{e6alt}, tricistronic vector of ERG^{wt} and $NRAS^{G12D}$, and ERG^{e6alt} and $NRAS^{G12D}$ for MIG; wild-type and IK6 (isoform derived from $IKZF1$ deletion of exon4-7) for MIR. These combinations resulted in twelve conditions. Transduced GFP and RFP-positive cells were obtained by fluorescence-activated cell sorting. For clonogenic assays, 10,000 cells were plated in triplicate in Methocult M3231 (Stem Cell Technologies, Inc., Vancouver, BC, Canada) with the appropriate factors (SCF, 100 ng/ml; IL-7, 20 ng/ml; FLT-3 ligand, 10 ng/ml) and colonies were scored 7 days later. For re-plating, 10,000 cells were cultured in identical conditions, with colonies counted on day 10–12. Colony identity was confirmed by morphological analysis through cytospin and Wright-Giemsa staining and by flow analysis for a panel of multi-lineage markers (Gr-1-PerCPCy5.5, Ter119-PE-Cy7, B220-eFluor605, CD3-APC, Mac1-Alexa700 and CD19-APC-Cy7) and specific B-cell lineage markers (CD43-PerCP-Cy5.5, IgM-PE-Cy7, and BP1-Alexa647).

Statistical analysis

Associations between categorical variables were examined using Fisher's exact test. Associations between $DUX4/ERG$ status with treatment outcome (event free survival and relapse) were performed as previously described[55–59]. Differences in survival between mice transplanted with ERG wild type or ERG alt transduced cells were compared using the Mantel Cox test. Analyses were performed using Prism v6.0 (Graphpad, La Jolla, CA), R (www.r-project.org)[60], SAS (SAS v9.1.2, SAS Institute, Cary, NC) and SPLUS (SPLUS 7.0, Insightful Corp., Palo Alto, CA) and StatXact (v 8.0.0, Cytel Inc, Cambridge, MA).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Jinghui Zhang[1], Kelly McCastlain[2], Hiroki Yoshihara[2], Beisi Xu[1], Yunchao Chang[2], Michelle L. Churchman[2], Gang Wu[1], Yongjin Li[1], Lei Wei[1,2], Ilaria Iacobucci[2], Yu Liu[1], Chunxu Qu[1], Ji Wen[1], Michael Edmonson[1], Debbie Payne-Turner[2], Kerstin B. Kaufmann[3], Shin-ichiro Takayanagi[3,4], Erno Wienholds[3], Esmé Waanders[2,5], Panagiotis Ntziachristos[6,*], Sofia Bakogianni[6], Jingjing Wang[6], Iannis Aifantis[6,7],

Kathryn G. Roberts[2], Jing Ma[2], Guangchun Song[2], John Easton[1], Heather L. Mulder[1], Xiang Chen[1], Scott Newman[1], Xiaotu Ma[1], Michael Rusch[1], Pankaj Gupta[1], Kristy Boggs[1], Bhavin Vadodaria[1], James Dalton[2], Yanling Liu[2], Marcus L Valentine[8], Li Ding[9], Charles Lu[9], Robert S. Fulton[9], Lucinda Fulton[9], Yashodhan Tabib[9], Kerri Ochoa[9], Meenakshi Devidas[10], Deqing Pei[11], Cheng Cheng[11], Jun Yang[12], William E. Evans[12], Mary V. Relling[12], Ching-Hon Pui[13], Sima Jeha[13], Richard C. Harvey[14], I-Ming L Chen[14], Cheryl L. Willman[14], Guido Marcucci, MD[15], Clara D. Bloomfield[16], Jessica Kohlschmidt[16], Krzysztof Mrózek[16], Elisabeth Paietta[17], Martin S. Tallman[18], Wendy Stock[19], Matthew C. Foster[20], Janis Racevskis[21], Jacob M. Rowe[25], Selina Luger[26], Steven M. Kornblau[24], Sheila A Shurtleff[2], Susana C. Raimondi[2], Elaine R. Mardis[9], Richard K. Wilson[9], John E. Dick[3], Stephen P Hunger[25], Mignon L. Loh[26], James R. Downing[2], and Charles G. Mullighan[2] **for the St Jude Children's Research Hospital – Washington University Pediatric Cancer Genome Project**

Affiliations

[1]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN [2]Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN [3]Princess Margaret Cancer Centre, University Health Network and Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5G 1L7, Canada [4]Oncology Research Laboratories, Kyowa Hakko Kirin Co., Ltd., Machida-shi, Tokyo, 194-8533, Japan [5]Department of Human Genetics, Radboud University Medical Center and Radboud Center for Molecular Life Sciences, Nijmegen, the Netherlands [6]Department of Pathology, New York University School of Medicine, New York, NY [7]Howard Hughes Medical Institute, New York, NY [8]Cytogenetics Core Facility, St. Jude Children's Research Hospital, Memphis, TN [9]McDonnell Genome Institute, Washington University, St Louis, MO [10]Department of Biostatistics, Colleges of Medicine, Public Health & Health Profession, University of Florida, Gainesville, FL [11]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN [12]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN [13]Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN [14]Department of Pathology, The Cancer Research and Treatment Center, University of New Mexico, Albuquerque, NM [15]Gehr Leukemia Center, City of Hope, Duarte, CA [16]Comprehensive Cancer Center, The Ohio State University, Columbus, OH [17]Cancer Center, Montefiore Medical Center North Division, Bronx, NY [18]Memorial Sloan-Kettering Cancer Center, New York, NY [19]University of Chicago Medical Center, Chicago, IL [20]Division of Hematology/Oncology, University of North Carolina, Chapel Hill, NC [21]Department of Medicine (Oncology), Albert Einstein College of Medicine, Montefiore Medical Center, New York, NY [22]Department of Hematology, Shaare Zedek Medical Center, Jerusalem, Israel [23]Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA [24]Section of Molecular Hematology and Therapy, Department of Leukemia, The University of Texas M.D. Anderson Cancer Center, Houston, TX [25]Department of Pediatrics, Children's Hospital of Philadelphia and the University of Pennsylvania

Perelman School of Medicine, Philadelphia, PA [26]Department of Pediatrics, Benioff Children's Hospital, University of California at San Francisco, San Francisco, CA

Acknowledgments

REFERENCES

1. Hunger SP, Mullighan CG. Acute Lymphoblastic Leukemia in Children. N Engl J Med. 2015; 373:1541–1552. [PubMed: 26465987]

2. Mullighan CG. Genomic characterization of childhood acute lymphoblastic leukemia. Semin Hematol. 2013; 50:314–324. [PubMed: 24246699]

3. Yeoh EJ, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell. 2002; 1:133–143. [PubMed: 12086872]

4. Harvey RC, 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. [PubMed: 20699438]

5. Reddy ES, Rao VN. erg, an ets-related gene, codes for sequence-specific transcriptional activators. Oncogene. 1991; 6:2285–2289. [PubMed: 1766675]

6. Mori H, et al. Chromosome translocations and covert leukemic clones are generated during normal fetal development. Proc Natl Acad Sci U S A. 2002; 99:8242–8247. [PubMed: 12048236]

7. Bartel FO, Higuchi T, Spyropoulos DD. Mouse models in the study of the Ets family of transcription factors. Oncogene. 2000; 19:6443–6454. [PubMed: 11175360]

8. Kruse EA, et al. Dual requirement for the ETS transcription factors Fli-1 and Erg in hematopoietic stem cells and the megakaryocyte lineage. Proc Natl Acad Sci U S A. 2009; 106:13814–13819. [PubMed: 19666492]

9. Loughran SJ, et al. The transcription factor Erg is essential for definitive hematopoiesis and the function of adult hematopoietic stem cells. Nat Immunol. 2008; 9:810–819. [PubMed: 18500345]

10. Salek-Ardakani S, et al. ERG is a megakaryocytic oncogene. Cancer Res. 2009; 69:4665–4673. [PubMed: 19487285]

11. Rainis L, et al. The proto-oncogene ERG in megakaryoblastic leukemias. Cancer Res. 2005; 65:7596–7602. [PubMed: 16140924]

12. Ng AP, et al. Trisomy of Erg is required for myeloproliferation in a mouse model of Down syndrome. Blood. 2010; 115:3966–3969. [PubMed: 20007548]

13. Tomlins SA, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005; 310:644–648. [PubMed: 16254181]

14. Prasad DD, Ouchida M, Lee L, Rao VN, Reddy ES. TLS/FUS fusion domain of TLS/FUS-erg chimeric protein resulting from the t(16;21) chromosomal translocation in human myeloid leukemia functions as a transcriptional activation domain. Oncogene. 1994; 9:3717–3729. [PubMed: 7970732]

15. Marcucci G, et al. Overexpression of the ETS-related gene, ERG, predicts a worse outcome in acute myeloid leukemia with normal karyotype: a Cancer and Leukemia Group B study. J Clin Oncol. 2005; 23:9234–9242. [PubMed: 16275934]

16. Rao VN, Papas TS, Reddy ES. erg, a human ets-related gene on chromosome 21: alternative splicing, polyadenylation, and translation. Science. 1987; 237:635–639. [PubMed: 3299708]

17. Dixit M, et al. DUX4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. Proc Natl Acad Sci U S A. 2007; 104:18157–18162. [PubMed: 17984056]

18. Italiano A, et al. High prevalence of CIC fusion with double-homeobox (DUX4) transcription factors in EWSR1-negative undifferentiated small blue round cell sarcomas. Genes Chromosomes Cancer. 2012; 51:207–218. [PubMed: 22072439]

19. Kawamura-Saito M, et al. Fusion between CIC and DUX4 up-regulates PEA3 family genes in Ewing-like sarcomas with t(4;19)(q35;q13) translocation. Hum Mol Genet. 2006; 15:2125–2137. [PubMed: 16717057]

20. Yasuda T, et al. Recurrent DUX4 fusions in B cell acute lymphoblastic leukemia of adolescents and young adults. Nat Genet. 2016; 48:569–574. [PubMed: 27019113]

21. van der Veer A, et al. IKZF1 status as a prognostic feature in BCR-ABL1-positive childhood ALL. Blood. 2014; 123:1691–1698. [PubMed: 24366361]

22. Mullighan CG, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N Engl J Med. 2009; 360:470–480. [PubMed: 19129520]

23. Young JM, et al. DUX4 binding to retroelements creates promoters that are active in FSHD muscle and testis. PLoS Genet. 2013; 9:e1003947. [PubMed: 24278031]

24. Zou J, et al. The oncogenic TLS-ERG fusion protein exerts different effects in hematopoietic cells and fibroblasts. Mol Cell Biol. 2005; 25:6235–6246. [PubMed: 15988032]

25. Carmichael CL, et al. Hematopoietic overexpression of the transcription factor Erg induces lymphoid and erythro-megakaryocytic leukemia. Proc Natl Acad Sci U S A. 2012; 109:15437–15442. [PubMed: 22936051]

26. Russell LJ, et al. Deregulated expression of cytokine receptor gene, CRLF2, is involved in lymphoid transformation in B-cell precursor acute lymphoblastic leukemia. Blood. 2009; 114:2688–2698. [PubMed: 19641190]

27. Mullighan CG, et al. Rearrangement of CRLF2 in B-progenitor- and Down syndrome-associated acute lymphoblastic leukemia. Nat Genet. 2009; 41:1243–126. [PubMed: 19838194]

28. Iacobucci I, et al. Truncating Erythropoietin Receptor Rearrangements in Acute Lymphoblastic Leukemia. Cancer Cell. 2016; 29:186–200. [PubMed: 26859458]

29. Russell LJ, et al. IGH@ translocations are prevalent in teenagers and young adults with acute lymphoblastic leukemia and are associated with a poor outcome. J Clin Oncol. 2014; 32:1453–1462. [PubMed: 24711557]

30. Zaliova M, et al. ERG deletion is associated with CD2 and attenuates the negative impact of IKZF1 deletion in childhood acute lymphoblastic leukemia. Leukemia. 2014; 28:182–185. [PubMed: 24072102]

31. Clappier E, et al. 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; 28:70–77. [PubMed: 24064621]

32. Mullighan CG, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature. 2007; 446:758–764. [PubMed: 17344859]

33. Lilljebjorn H, et al. Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. Nat Commun. 2016; 7:11790. [PubMed: 27265895]

METHODS-ONLY REFERENCES

34. Zhou X, et al. Exploring genomic alteration in pediatric cancer using ProteinPaint. Nat Genet. 2016; 48:4–6. [PubMed: 26711108]

35. Holmfeldt L, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nat Genet. 2013; 45:242–252. [PubMed: 23334668]

36. Zhang J, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature. 2012; 481:157–163. [PubMed: 22237106]

37. Zhou X, et al. Exploring genomic alteration in pediatric cancer using ProteinPaint. Nat Genet. 2015; 48:4–6.

38. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31:166–169. [PubMed: 25260700]

39. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004; 3 Article3.

40. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995; 57:289–300.

41. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013; 10:1213–1218. [PubMed: 24097267]

42. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]

43. Shultz LD, et al. Human lymphoid and myeloid cell development in NOD/LtSz-scid IL2R gamma null mice engrafted with mobilized human hemopoietic stem cells. J Immunol. 2005; 174:6477–6489. [PubMed: 15879151]

44. Kneissl S, et al. Measles virus glycoprotein-based lentiviral targeting vectors that avoid neutralizing antibodies. PLoS One. 2012; 7:e46667. [PubMed: 23071609]

45. Churchman ML, et al. Efficacy of Retinoids in IKZF1-Mutated BCR-ABL1 Acute Lymphoblastic Leukemia. Cancer Cell. 2015; 28:343–356. [PubMed: 26321221]

46. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

47. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

48. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008; 26:1351–1359. [PubMed: 19029915]

49. Robinson JT, et al. Integrative genomics viewer. Nat Biotechnol. 2011; 29:24–26. [PubMed: 21221095]

50. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics. 1998; 14:48–54. [PubMed: 9520501]

51. Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37:W202–W208. [PubMed: 19458158]

52. Mathelier A, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016; 44:D110–D115. [PubMed: 26531826]

53. Mulligan CG, et al. BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. Nature. 2008; 453:110–114. [PubMed: 18408710]

54. Kamijo T, et al. Tumor suppression at the mouse INK4a locus mediated by the alternative reading frame product p19ARF. Cell. 1997; 91:649–659. [PubMed: 9393858]

55. Mulligan CG, et al. Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia. N Engl J Med. 2009; 360:470–480. [PubMed: 19129520]

56. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep. 1966; 50:163–170. [PubMed: 5910392]

57. Peto R, et al. Design analysis of randomized clinical trials requiring prolonged observation of each patient II. analysis and examples. Br J Cancer. 1977; 35:1–39. [PubMed: 831755]

58. Gray RJ. A class of *K*-sample tests for comparing the cumulative incidence of a competing risk. Annals Statistics. 1988; 16:1141–1154.

59. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. J Am Stat Assoc. 1999; 94:496–509.

60. R Development Core Team. R Foundation for Statistical Computing. Austria: Vienna; 2009. R: A language and environment for statistical computing.
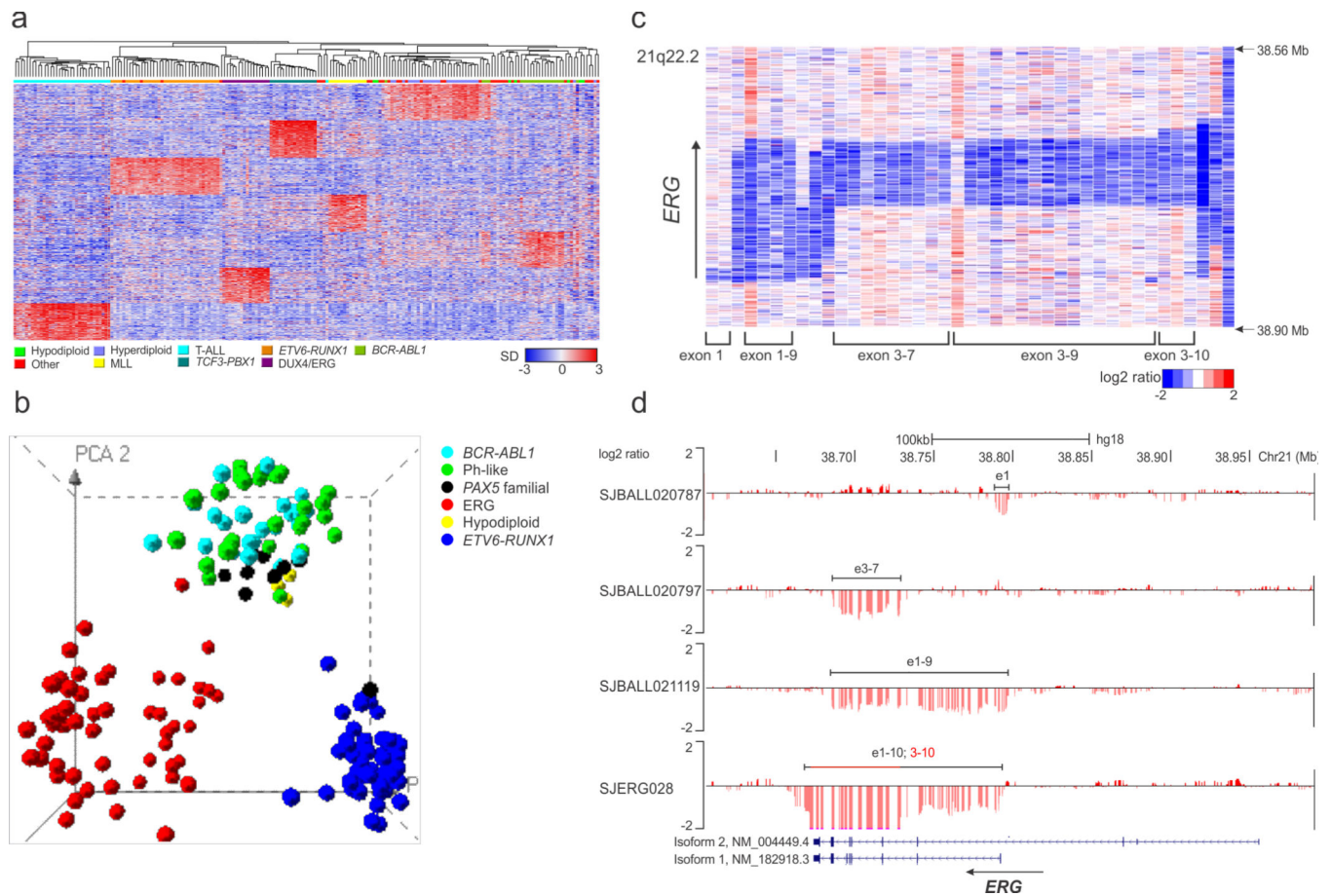
Figure 1. Gene expression profile and *ERG* deletions in *DUX4/ERG* ALL

a, Hierarchical clustering of the top 100 Affymetrix U133A probe sets upregulated in each subtype of 199 B- and T-lineage ALL cases, including *DUX4/ERG* ALL. **b**, Principal component analysis of FPKM gene expression data derived from poly-A RNA sequencing identifying *DUX4/ERG* ALL **c**, SNP microarray data of cases with *ERG* deletions. Data for each case is shown as a column, and shown in a $\log_2$ ratio scale where deletions are blue. The position of the *ERG* locus is shown as an arrow on the left of the panel, and genomic coordinates shown in megabases on the right. Cases are ordered from left to right according to the extent of the deletion, with focal exon 1 deletions on the left, the common exons 3–7 and 3–9 deletions in the middle of the panel, and whole gene deletion on the right. **d**, Representative DNA copy number data of four cases with *ERG* deletion also shown in $\log_2$ ratio scale, with probe level data shown, deletion being below the *x* axis. The extent of deletion in each case is shown by horizontal bars.
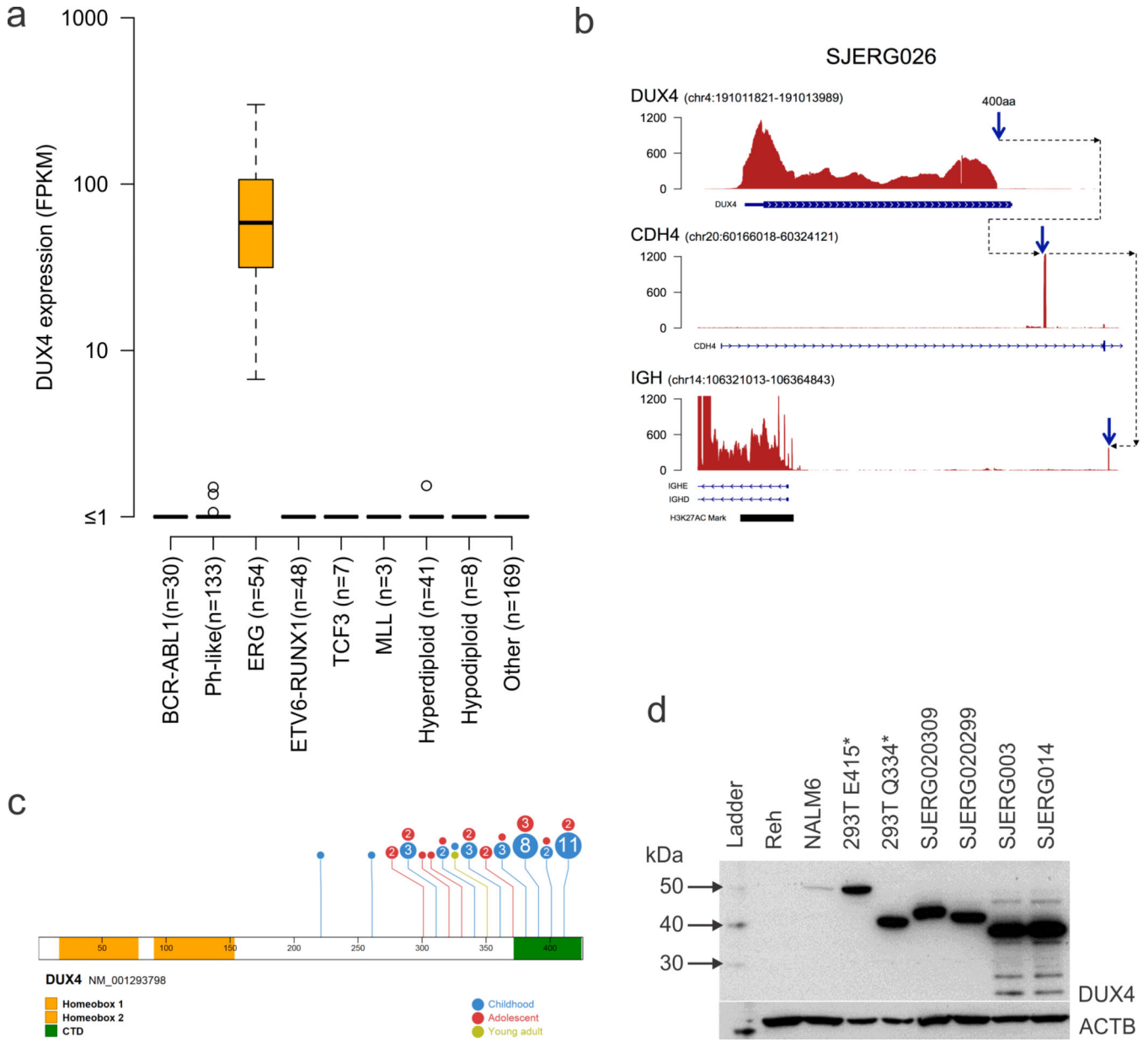
Figure 2. Rearrangement of *DUX4*

a, deregulation of *DUX4* is observed exclusively in the ALL cases with *ERG* deregulation. Gene expression data shown as fragments per kilobase of mapped reads (FPKM) from RNA-sequencing in box plots, with ALL cases grouped by subtype on the *x* axis. Horizontal lines show median, and the boxes, interquartile range. **b**, an example of the commonly complex rearrangements of *IGH* and *DUX4*. Each third of the panel shows RNA-seq gene expression data for three loci involved in the rearrangement. The top shows increased expression at *DUX4*. The dotted line shows the breakpoint of *DUX4* that is juxtaposed to a small segment of *CDH4* shown in the middle panel, which in turn is rearranged to *IGH*, shown in the lower panel. **c**, schematic showing the location of the breakpoints in *DUX4*, stratified by age group. **d**, Immunoblotting of cell line and primary leukemic sample lysates showing proteins
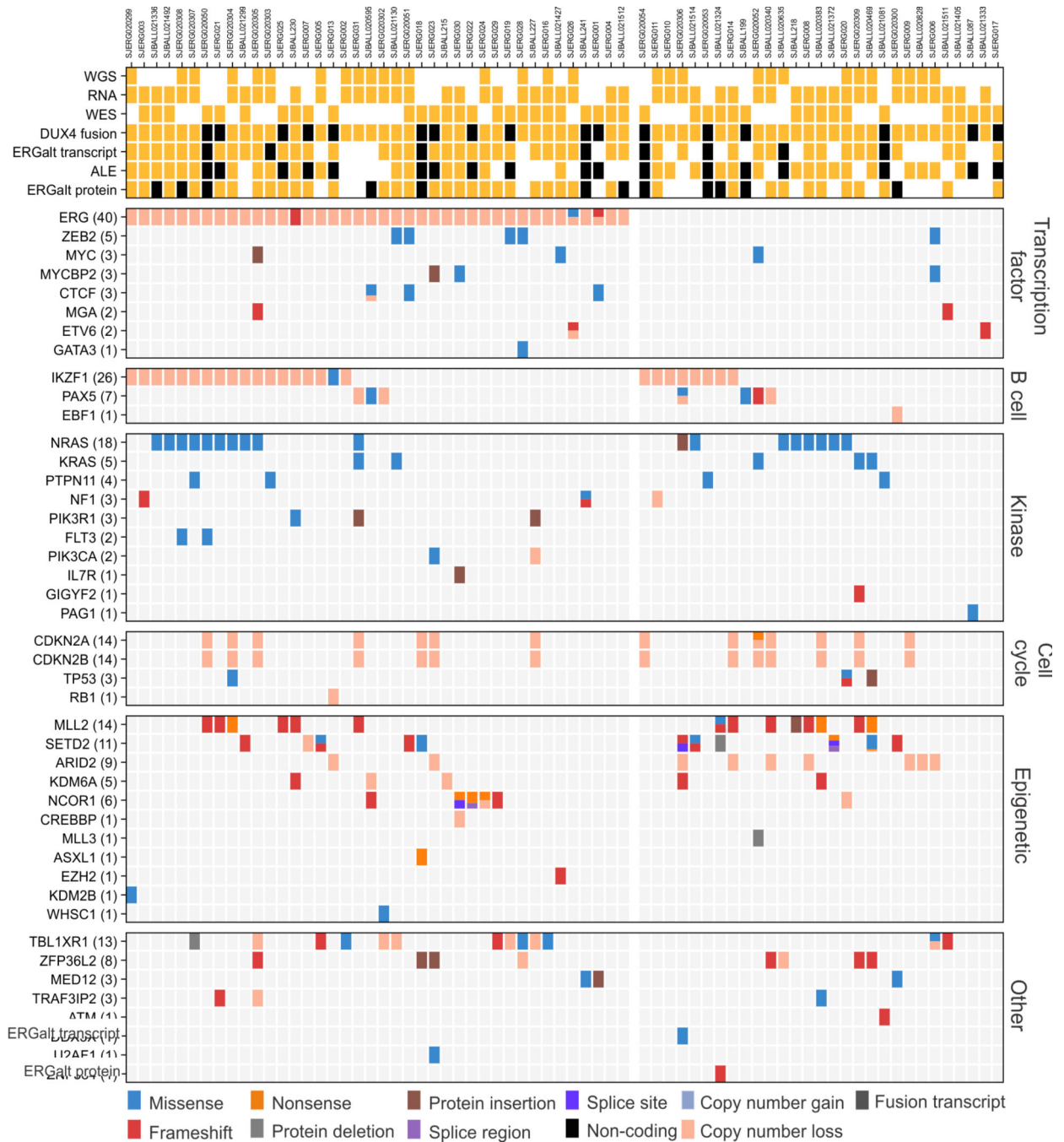
of variable size corresponding to DUX4 C-terminal truncation and/or appending with amino acid residues encoded by read through into the IGH locus.

Figure 3. Structural and sequence alterations in *DUX4/ERG* ALL

Heatmap showing genomic data for *DUX4/ERG* ALL cases, each of which is represented in a column and genes are grouped by functional pathway. The colors for each type of genetic alterations are shown at the bottom of the figure. The genomic profiling performed, and presence/absence of ERGalt, ALE and DUX4 rearrangement are shown at the top of the figure, where yellow represents assay performed (or alteration present), white, not performed or absent, and gray, data not available.
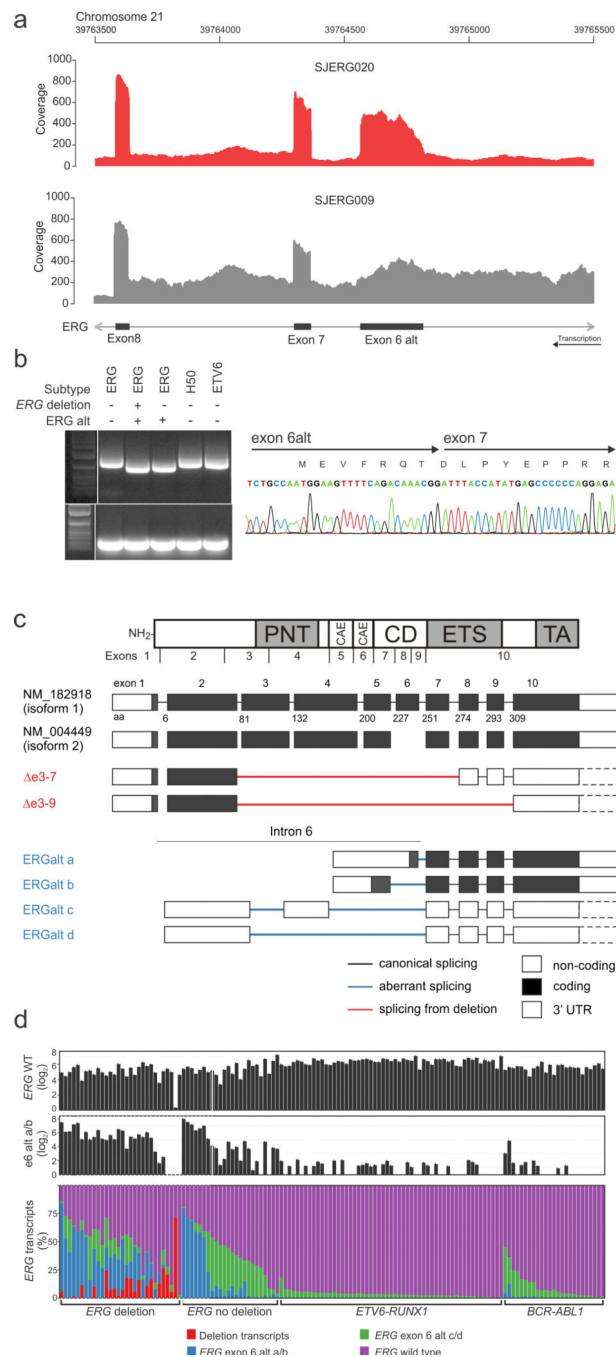
Figure 4. Expression of ERGalt in DUX4/ERG ALL

a, read depth from mRNA-sequencing of an ERG ALL case with expression of exon 6 alt (red) and a case lacking ERG alt expression (gray). **b**, RT-PCR with PCR primers specific to exon 6 alt and exon 10, showing amplification of a larger amplicon in ALL cases lacking ERG alt, with amplification of the intervening intronic sequence between ERGalt and exon 7, and a smaller amplicon in ERGalt positive cases arising from splicing from exon 6 alt to exon 7, shown in a representative electropherogram. This transcript results in a novel N-terminus of ERG encoded by exon 6 alt comprising 7 residues spliced in frame to exons 7–

10. **c**, structure of the canonical and ERG ALL-associated ERG transcripts and isoforms. **d**, relative abundance of ERG transcripts in ERG ALL and other B-ALL subtypes. Each column represents a case. The top panel represents expression of wild type (WT) ERG, and the middle panel expression of ERGalt, and the less abundant ERGalt isoforms. The lower panel shows the cumulative abundance of all isoforms as a proportion within each case. Transcripts arising from splicing across *ERG* deletion are low abundance in most cases with *ERG* deletion (with the exception of one case with biallelic deletion. Non-coding ERGalt isoforms are more common in ERG ALL cases lacking ERG deletion.
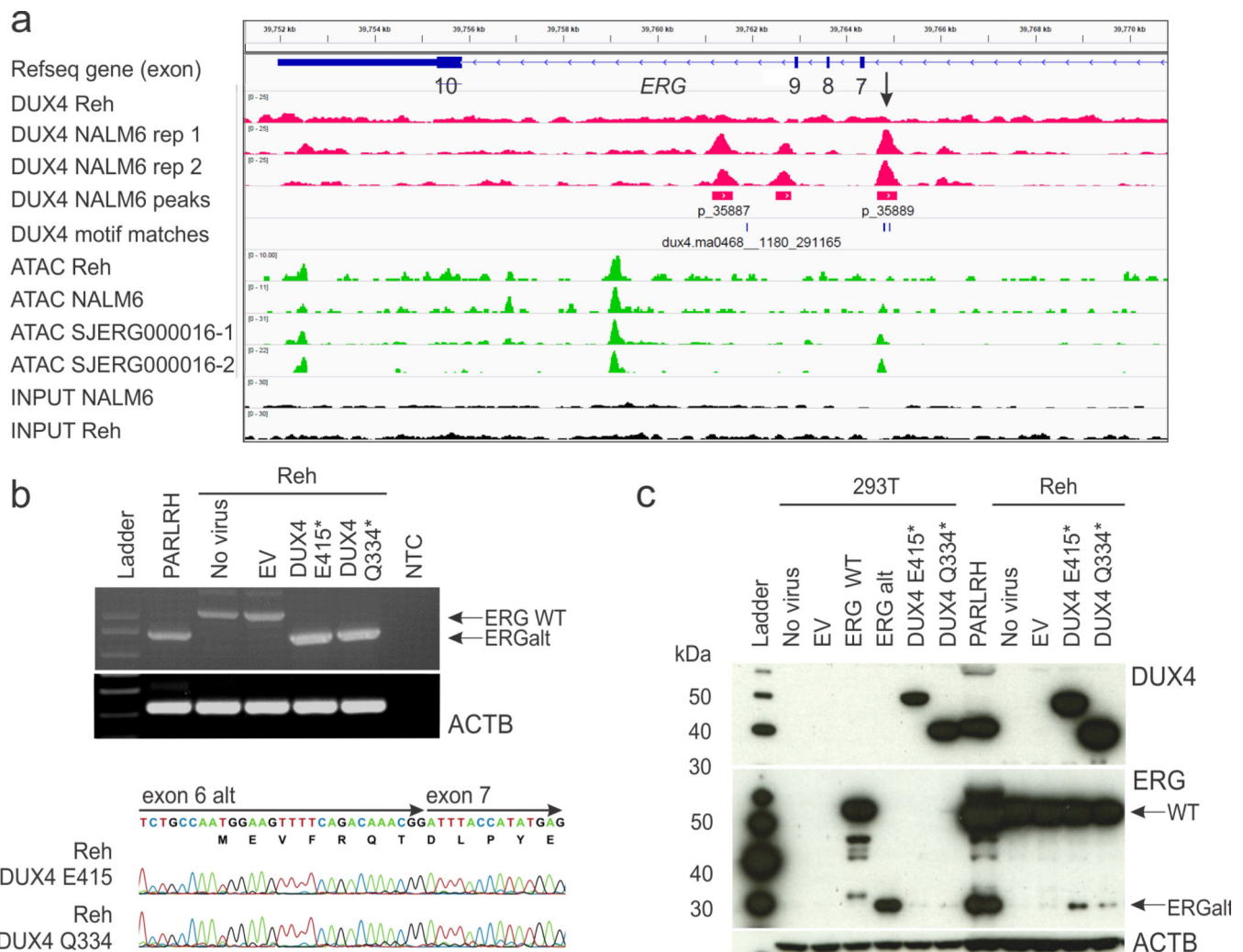
Figure 5. DUX4 induces deregulation of ERG

a, DUX4 ChIP-seq and ATAC-sequencing data of Reh, NALM6 and *DUX4/ERG* ALL xenograft showing DUX4 binding at *ERG* exon 6 alt (arrowed) in NALM6 but not in Reh. ATAC sequencing showing open chromatin at this peak of DUX4 binding at *ERG* exon 6 alt in the DUX4/ERG xenograft sample and NALM6 but not in Reh. ATAC-seq data is provided in Supplementary Figure 19. **b–c**, expression of two truncated DUX4 alleles, but not empty vector, results in expression of ERGalt by RT-PCR (b) and immunoblotting (c). **b**, the top panel shows RT-PCR for *ERG* using primers specific for exon 6 alt and exon 7. In untransduced or empty vector transduced Reh cells, a larger amplicon is observed that incorporates the intervening intronic transcript (lanes 3 and 4). In *DUX4/ERG* ALL cells (PARLRH, lane 2) and *DUX4*-transduced Reh cells (lanes 5 and 6), ERGalt is amplified, with in-frame splicing from exon 6 alt to exon 7 (lower part of panel). **c**, Immunoblotting showing expression of ERGalt in cells transduced with DUX4. The top part of the panel shows immunoblotting with an N-terminus DUX4-specific antibody; the middle part blotting with a C-terminus specific ERG antibody, and the lower part, actin control. DUX4 alleles induce expression of ERGalt in HEK293T cells (lanes 6 and 7) and Reh cells (lanes
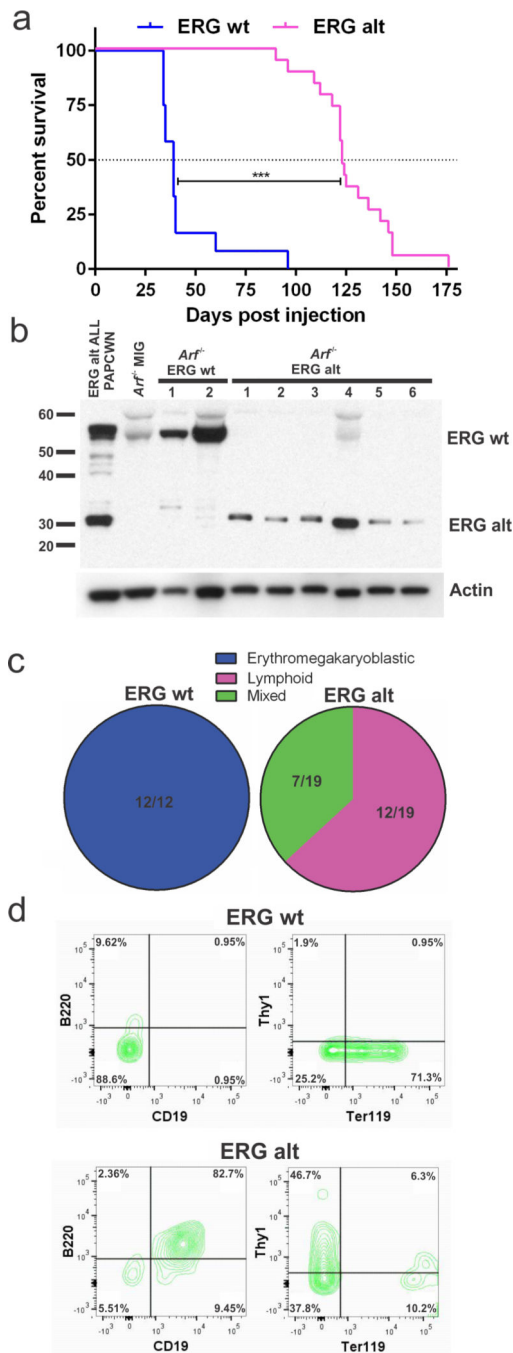
11 and 12). Lane 5 shows 293T cells transfected with ERGalt virus as a positive control for ERGalt (and negative for DUX4); lane 8 shows DUX4/ERG patient sample PARLRH positive for DUX4 and ERGalt.

Figure 6. Expression of ERGalt induces ALL

a, Transplantation of lineage negative Arf$^{-/-}$ progenitors induces erythromegakaryoblastic leukemia, and ERGalt expressing marrow a lymphoid progenitor, biphenotypic or B-lymphoid leukemia. *** $P < 0.0001$ (Mantel-Cox); n= 12 ERG WT, 19 ERGalt mice from 2 independent experiments. **b**, immunoblotting with a C-terminus specific ERG antibody for ERG ALL samples and splenocytes of mouse leukemias, showing expression of ERG alt in human ERG ALL and mouse tumors induced by expression of ERG alt. **c**, proportion of leukemias displaying erythromegakaryoblastic, lymphoid, or mixed immunophenotype

(mixed = lymphoid and myeloid subpopulations). **d**, representative immunophenotyping of ERG-induced tumors showing expression of the erythroid marker Ter119 in an ERG WT induced tumor, and coexpression of B220 and CD19 in ERGalt induced leukemia. Cells were gated on GFP expression.