SUPPLEMENTARY APPENDIX

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

Marketa Zaliova,[1,2,3] Eliska Potuckova,[1,2] Lenka Hovorkova,[1,2] Alena Musilova,[1,2] Lucie Winkowska,[1,2] Karel Fiser,[1,2] Jan Stuchly,[1,2] Ester Mejstrikova,[1,2,3] Julia Starkova,[1,2] Jan Zuna,[1,2,3] Jan Stary[2,3] and Jan Trka[1,2,3]

[1]CLIP - Childhood Leukaemia Investigation Prague; [2]Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University, Prague and [3]University Hospital Motol, Prague, Czech Republic

 doi:10.3324/haematol.2018.204487

Received: August 13, 2018.

Accepted: January 7, 2019.

Pre-published: January 10, 2019.

Correspondence: *MARKETA ZALIOVA* - marketa.zaliova@lfmotol.cuni.cz

SUPPLEMENTAL APPENDIX

ERG deletions in childhood acute lymphoblastic leukemia with DUX4 rearrangements are mostly polyclonal, prognostically relevant and their detection rate strongly depends on method's sensitivity

Marketa Zaliova[1,2,3], Eliska Potuckova[1,2], Lenka Hovorkova[1,2], Alena Musilova[1,2], Lucie Winkowska[1,2], Karel Fiser[1,2], Jan Stuchly[1,2], Ester Mejstrikova[1,2,3], Julia Starkova[1,2], Jan Zuna[1,2,3], Jan Stary[1,3] and Jan Trka[1,2,3]

[1] CLIP - Childhood Leukaemia Investigation Prague
[2] Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University, Prague, Czech Republic
[3] University Hospital Motol, Prague, Czech Republic

Contents

Supplemental Methods

Sample processing

Leukemic cells were sorted using a fluorescence-assisted-cell-sorter from the samples with < 70% of leukemic blasts, otherwise bulk samples were used. Mononuclear cells from the bulk diagnostic and remission samples were isolated using the Ficoll-Paque density centrifugation method (Pharmacia, Uppsala, Sweden). Total DNA and RNA were isolated from mononuclear cells or from sorted blasts as part of routine diagnostics. The integrity of RNA was analyzed by chip electrophoresis using the RNA 6000 Nano Kit on the Agilent 2100 Bioanalyzer (Agilent Technologies, Carlsbad, CA, USA). The concentration of the nucleic acids was determined by spectrophotometry using the NanoDrop 2000 or by fluorometry using the Qubit® dsDNA BR Assay Kit on the Qubit® 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA).

Single nucleotide polymorphism (SNP) array

Copy number aberrations (CNA) and regions of uniparental disomy (UPD) were identified using HumanOmni Express BeadChip (Illumina, San Diego, CA, USA) or CytoScan HD arrays (Affymetrix, Santa Clara, CA, USA).

- *HumanOmni Express BeadChip (63 patients)*

DNA labeling and hybridization were performed according to the Infinium HD assay Ultra protocol from Illumina. The GenomeStudio software v2011.1 (Illumina) was used for genotype calling and quality control. Copy number variations (CNV) and UPD were called using the CNV Partition 2.4.4 algorithm plug-in within the GenomeStudio software. The resulting data (Log R ratio corresponding to copy number and B allele frequency corresponding to SNP genotype) were visually inspected in the Illumina Chromosome Browser.

- *CytoScan HD arrays (41 patients)*

Analysis was performed as a service in the Laboratory for Molecular Biology and Tumor Cytogenetics at the Department of Internal Medicine of Hospital Barmherzige Schwestern (Linz, Austria). The Chromosome Analysis Suite software (Affymetrix) was used for quality control, genotype calling, CNV/UPD identification and data visualization.

Results from both platforms were manually curated. Deletions corresponding to somatic rearrangements of the immunoglobulin and T-cell receptor gene loci, germline CNV/UPD (present in remission samples) and common population variations were excluded.

Whole transcriptome sequencing

Whole transcriptome sequencing (RNAseq) was performed in 112/118 patients. Sequencing libraries were prepared using Agilent SureSelect mRNA Strand Specific (RNAseq) kits according to the manufacturer's instructions (Agilent, Technologies, USA). High-throughput sequencing (2x75 cycles) was performed on HiSeq2500 or NextSeq500 using TruSeq Rapid SBS and PE Cluster kits and High Output Kit, respectively, according to the manufacturer's instructions (Illumina).

Gene expression profiling on microarrays

Gene expression profiling (GEP) was performed in 6 patients who were not analyzed by RNAseq and in additional 72 patients who were analyzed also by RNAseq. GEP was performed on HumanHT-12 v4 Expression BeadChip (Illumina). The raw data were preprocessed using GenomeStudio software (Illumina) and further processed using R interface (version 3.2.2). Briefly, the transcription profiles were background corrected, quantile normalized and variance stabilized using log2-transformation.

Hierarchical clustering

Patients were clustered hierarchically (Euclidean distance linkage and ward.D method) based on the normalized expression of genes belonging to the DUX4r ALL specific gene set, using data either from RNAseq or from GEP. The *DUX4*r ALL specific gene set was extracted from the analysis of 23 *DUX4*r/*ERG* and 125 non-*DUX4*r/*ERG* ALL cases performed by Zhang et al.[1] The set included top 150 up- and top 150 down-regulated genes (ranked by fold-change) in *DUX4*r-ALL compared to non-*DUX4*r ALL. Genes are listed in Supplemental Table 4.

Identification of the reads corresponding to the DUX4 gene rearrangement at the RNA level

Mapped reads were visualized in Integrative Genome Viewer[2], and the following reads supporting the presence of DUX4 gene rearrangements were manually searched at both D4Z4 repeat regions (4q,10q):

- Reads with unmapped parts ("softclipped" sequences) that match the IGH gene reference (or to some region outside of chromosomes 4 and 10)
- Reads with mates mapped to the IGH gene locus (or to some region outside of chromosomes 4 and 10)

Similarly, the region of a potential fusion partner (identified in the previous step) was then inspected for the presence of unmapped read parts matching (or read mates mapped) to the DUX4 reference sequence (or D4Z4 regions).

Identification of the reads corresponding to ERGalt at the transcriptomic level

Reads containing the sequences specific for *ERG*alt were counted and normalized to library sizes (to correct for the uneven sequencing depth, size factor was computed by DeSeq2[3]). The following 16bp long sequences matching the junctions of alternative first exons (that are specific for *ERG*alt and are not involved in wild-type *ERG* isoforms[1]) to common second exon of ERGalt were used: GACAAACGGATTTACCA (ERGalt_a), ACTAGATTATTTACCA (ERGalt_b). These sequences (at 100% of their lengths) do not match any other human transcript according to BLAT and BLAST tools (https://genome.ucsc.edu/cgi-bin/hgBlat, https://blast.ncbi.nlm.nih.gov/Blast.cgi).

IGH-DUX4 capture sequencing and analysis of the DUX4 gene rearrangement at the genomic level

To analyze presence of *IGH-DUX4* at the genomic level, capture sequencing was performed in 17 patients. Sequencing libraries were prepared form 50ng genomic DNA using SureSelectXT Custom Target Enrichment Probes and SureSelectQXT Reagent Kit NSQ (Agilent) according to the manufacturer's instructions. High-throughput sequencing (2x150 cycles) was performed on NextSeq500 using Mid Output Kit according to the manufacturer's instructions (Illumina). Following genomic regions were targeted by custom capture probes: chr4:190989000-191015000, chr10:135479000-135499000, chr14:106310000-106395000. Read pairs were aligned to the human genome reference (hg19) using BWA[4]. Mapped reads were visualized in Integrative Genome Viewer. The *IGH-DUX4* fusion was called based on:

(i) The presence of reads mapped to *IGH* with mates mapped within one or both D4Z4 repeat regions on 4q and 10q (automatic read search),

(ii) The presence of reads partially mapped to *IGH* whose unmapped parts matched the *DUX4* sequence or other sequence within D4Z4 repeat regions (manual read search, the BLAT tool of UCSC Genome Browser was used for the alignment of unmapped read parts to the human genome).

Analysis of data from amplicon sequencing

Reads from amplicon sequencing (AmpliSeq) were successively mapped to a custom reference as described in Materials and methods. For the mapping, the default setting of the bwa aligner[4] was modified as follows: gap open penalty (-O) 6, gap extension penalty (-E) 4, clipping penalty (-L) 3.

In addition to investigation of the presence/absence of ERGdel, the ERGdel repertoire (set of distinct coexisting ERGdel subclones), was analyzed in AmpliSeq-positive patients who were sequenced at higher target coverage setting (n=24). Set of distinct ERGdel alleles identified via bioinformatic algorithm (and defined by the position of last non-deleted 5' nucleotide, N-segment sequence, type of utilized 3' breakpoint site cluster and position of the first non-deleted 3' nucleotide) was manually curated in order to exclude contaminations and sequencing errors. Ultra deep sequencing of plasmids containing unique ERGdel sequences was used to analyze repertoire of sequencing errors inherent to the used sequencing platform.

ERGdel clones reported as distinct were joined into a single IntERGdel clone (defined by the sequence of the most abundant clone) if the differences in their sequences were considered as likely sequencing artifacts. These were:

- Shortenings and extensions of the homo-di-/oligo-mers
- Single base duplications, deletions, insertions and substitutions (1-2 per each of the following sequence segments: the last 10 bases of the 3' segment, N-segment, the first 10 bases of the 5' segment)

In each patient, a maximum of 10 different most abundant ERGdel clones for each of the five possible 3' breakpoint site clusters were curated and counted within final ERGdel repertoire in table 2. When the same ERGdel sequence was identified in the final repertoire of more than one sample, it was considered as a potential contamination and excluded from the IntERGdel repertoire of the sample in which it was covered by a lower amount of reads compared to another sample from the same (or preceding) sequencing run.

Supplemental Results

Identification of DUX4r ALL subtype within B-other ALL

Classification into *DUX4*r ALL subtype was primarily done by hierarchical clustering analysis (HCA) based on gene expression profiles and secondarily confirmed by the presence of *IGH-DUX4* (*ERG-DUX4*) transcripts and/or genes.

HCA based on expression of the previously published *DUX4*r ALL signature genes[1] clearly separated patients into 2 clusters using gene expression data both from RNAseq and from microarray (Supplemental Figures 1 and 2). One of two clusters (in both analyses) contained (with a single exception in RNAseq-based HCA) all *ERG*del-positive patients, and ALL cases (n=50) within this cluster were classified as *DUX4*r ALL; the remaining cases (n=68) were classified as non-*DUX4*r ALL. Results of these 2 clustering analyses were fully concordant; all 68 ALL cases with both types of data available were classified concordantly. As described previously, *DUX4* expression differed gradually within *DUX4r* ALL, however, all *DUX4*r ALL patients had higher *DUX*4 expression levels compared to non-*DUX4*r ALL patients (Supplemental Figure 3).

Targeted analysis of RNAseq data revealed presence of fusion transcripts supporting the presence of *IGH-DUX4* or *ERG-DUX4* rearrangements in 42/44 and 1/44 *DUX4*r ALL patients with available RNAseq data, respectively. No reads supporting the presence of *DUX4* rearrangement were found at the RNA level in the remaining 1 *DUX4*r ALL patient and in all non-*DUX4*r ALL patients. Targeted sequencing of *IGH* and *DUX4* loci at the genomic level via capture sequencing was performed in selected cases to further verify the classification results. The *IGH-DUX4* fusion was identified in all six *DUX4*r ALL cases without RNAseq data (who were classified as *DUX4*r based on HCA of gene expression data from microarrays) as well as in a single *DUX4*r ALL patient who lacked fusion-supporting reads at the RNA level. Next, we analyzed 6 *DUX4*r ALL patients with varying levels of *DUX4* expression including the patient with the highest level and 3 of 4 patients with the lowest levels (the fourth of these patients had *ERG-DUX4*). In accordance with the presence of *IGH-DUX4* at the RNA level, all patients were positive for *IGH-DUX4* also by capture sequencing. Finally, capture sequencing confirmed the absence of *IGH-DUX4* in 4 non-*DUX4*r ALL patients with the highest *DUX4* expression level. Altogether, these data show that via HCA we correctly identified *DUX4*r ALL. The information about presence/absence of *DUX4*r at RNA and genomic levels and *DUX4* expression levels is included in Supplementary Table 2.

Expression of ERGalt in DUX4r and non-DUX4r ALL

Zhang et al. showed that the *DUX4* rearrangement is accompanied by the expression of several novel *ERG* isoforms, two of which - *ERG*alt a and b - are protein-coding (hereafter termed *ERG*alt)[1]. These *ERG*alt transcripts initiate in an intronic region that is lost on the allele with Int*ERG*del and thus they can only be expressed from the second *ERG* allele in Int*ERG*del-positive cases[1]. Of note, Zhang et al. showed, that *ERG*alt transcripts are expressed also in minority of non-*DUX4*r ALL cases, although, the high expression levels of *ERG*alt were only observed in *DUX4*r ALL[1]. Importantly, they also showed that a minor proportion of *DUX4*r ALL cases did not express *ERG*alt (Supplementary Tables 11 and 12 in Zhang et al.)[1].

We analyzed the presence of *ERG*alt specific junction reads in RNAseq data. Such reads (≥1 read) were present in 41/44 and 61/68 *DUX4*r and non-*DUX4*r ALL, respectively. In accordance with Zhang

et al., high numbers of *ERG*alt specific junction reads (mirroring the high *ERG*alt expression level) were present only in *DUX4*r ALL (Supplemental Table 2, Supplemental Figure 4).

References

1. Zhang J, McCastlain K, Yoshihara H, et al. Deregulation of DUX4 and ERG in acute lymphoblastic leukemia. *Nat Genet*. 2016;**48**(12):1481-1489.
2. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;**29**(1):24-26.
3. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;**15**(12):550.
4. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;**26**(5):589-595.

Supplemental Tables

Supplemental Table 1. Primers for the multiplex PCR reaction used to prepare deep-sequencing libraries

Primer	Primer sequence (5' to 3')
	Deletion spanning amplicon (varying amplicon length)
A	CCATCTCATCCCTGCGTGTCTCCGACTCAG***XXXXXXXXXX***GAT**TCAGCTTCGGTGGTCCACAG**
trP1-BP1	CCTCTCTATGGGCAGTCGGTGAT**CTATCCTGAACATTGCTGCCAG**
trP1-BP2	CCTCTCTATGGGCAGTCGGTGAT**TGTAGATTCTCTTGCAGGGATGTACACTG**
trP1-BP3	CCTCTCTATGGGCAGTCGGTGAT**CCCTATGTTGAAATCTTAACCCGCAG**
trP1-BP4	CCTCTCTATGGGCAGTCGGTGAT**GCGGCTACTTGTTGGTCCAAGAA**
trP1-BP5	CCTCTCTATGGGCAGTCGGTGAT**GTCTAACTCAGAAGCATCTCACGGTAAGG**
	Control amplicon 1 (gene specific amplicon length 133 bp)
A	CCATCTCATCCCTGCGTGTCTCCGACTCAG***XXXXXXXXXX***GAT**GATAAGATGCCCCATTGCCCAG**
trP1	CCTCTCTATGGGCAGTCGGTGAT**AAGGGAGAAGAGAGTAGACAGTGATGGAG**
	Control amplicon 2 (gene specific amplicon length 271 bp)
A	CCATCTCATCCCTGCGTGTCTCCGACTCAG***XXXXXXXXXX***GAT**GGGAGGTGGCTGAGAGTTCACTCTC**
trP1	CCTCTCTATGGGCAGTCGGTGAT**CCAGTCCCAGAAGTCACACACTGTCAC**

Gene specific parts of the primers are in bold. "A" primers involve different, unique, 10-nucleotides long indexes (represented by "X" in bold italics).
bp – base pairs; BP – breakpoint site

Supplemental Table 2 „Patient characteristics and overview of the results" and Supplemental Table 3 „IntERGdel repertoire" are provided as separate excel files.

Supplemental Table 4. DUX4r ALL specific gene set

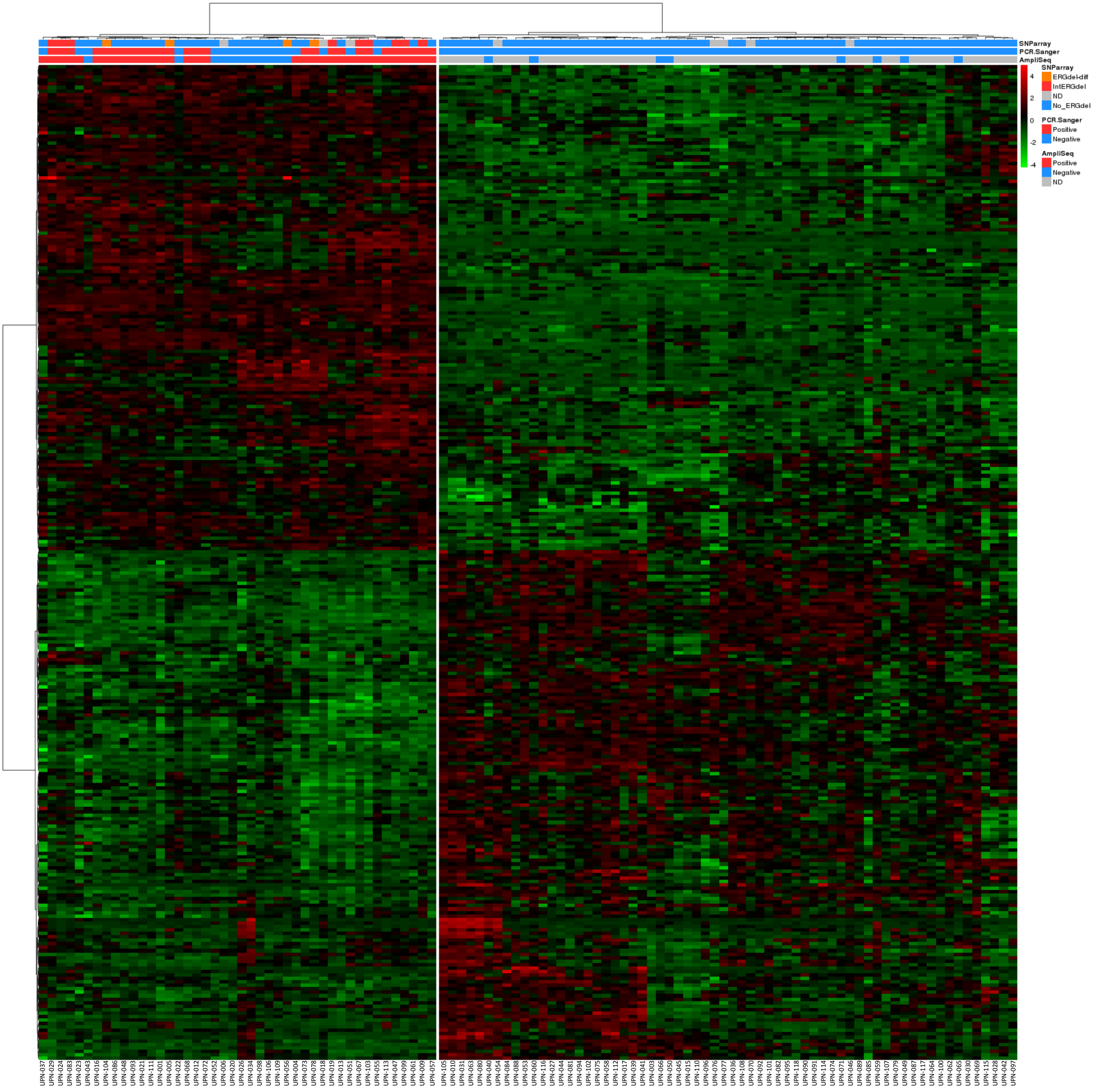
Upregulated in DUX4r ALL			
ABCB4	*ESAM*	*LRRC3B*	*RIMS3*
ADAM33	*F2RL3*	*MAST4*	*RIN2*
ADAMTS15	*FAM125B*	*MCAM*	*SALL4*
ADAMTS7	*FAM129C*	*MKRN3*	*SATB2*
ADAMTS9	*FJX1*	*MMP14*	*SCN2A*
AFF3	*FLJ35946*	*MPPED2*	*SEMA6D*
AGAP1	*FOXL1*	*MTSS1L*	*SERINC2*
AKT3	*GATA3*	*MTX2*	*SLC25A23*
ANGPT2	*GATM*	*N4BP3*	*SLC29A2*
ANKRD30B	*GCNT1*	*NDRG2*	*SNTB1*
B3GALNT1	*GLDC*	*NDRG4*	*SOX13*
B3GNT7	*GOLIM4*	*NFATC4*	*SOX7*
C1orf186	*GPR155*	*NFIL3*	*SPIB*
C1orf226	*GPSM1*	*NXN*	*SPP1*
C21orf91	*HEY2*	*OSBPL1A*	*STAP1*
C6orf25	*ID4*	*PAM*	*TACC2*
CCDC162P	*IGF2*	*PCBP4*	*TBX2*
CD200R1	*IKZF2*	*PCDH17*	*TCERG1L*
CD34	*ISM1*	*PCDH9*	*TDRD9*
CD38	*ITGA6*	*PDE8B*	*TLE2*
CD84	*ITM2A*	*PDGFRA*	*TMCC3*
CDH11	*KCNG1*	*PDLIM1*	*TMEM121*
CHRNB4	*KCNQ5*	*PGBD5*	*TMEM17*
CHST2	*KLHL13*	*PHACTR3*	*TMEM181*
CHST7	*KLRK1*	*PHYH*	*TOX3*
CLEC12A	*LDLRAD3*	*PLTP*	*TSKU*
CLEC12B	*LGR5*	*PMP22*	*TXNRD3*
CLTC	*LGR6*	*PPFIA4*	*UCK2*
CNR1	*LHFPL2*	*PPM1H*	*VLDLR*
CSMD1	*LIMCH1*	*PTGFRN*	*ZBTB46*
CXXC5	*LNX1*	*PTPRM*	
CYP46A1	*LOC100499467*	*PTPRS*	
DDIT4L	*LOC100506013*	*PVRL1*	
DDN	*LOC100507351*	*RAB20*	
DHX32	*LOC149086*	*RBFOX2*	
DLL3	*LOC284551*	*RGMB*	
DUX2	*LOC728989*	*RGS1*	
DUX4L4	*LOXL4*	*RGS16*	
EHD4	*LRRC2*	*RGS9*	
EPHA4	*LRRC28*	*RHOBTB1*	

Downregulated in DUX4r ALL			
TSPAN7	*CD99*	*NR3C2*	*MYO10*
ECM1	*BIRC7*	*AHR*	*ARHGAP29*
KHDRBS3	*CSF3R*	*ANO1*	*SLC35E3*
NT5E	*PECAM1*	*NTNG2*	*SULF2*
TMEM236	*TIAM2*	*ZC3H12D*	*MTSS1*
CYB5R2	*IRAK3*	*CHST15*	*MRVI1*
CTGF	*LRRC70*	*TSPAN15*	*TMC8*
FAM69C	*GRK5*	*TNF*	*AIF1*
GPR56	*CLIC5*	*BZRAP1*	*ALDH1A1*
GPR110	*TBC1D9*	*NEURL1B*	*ELFN2*
CMTM2	*PCLO*	*C5orf62*	*DHRS3*
OVCH2	*FARP1*	*RAB11FIP5*	*FBXW7*
LOC100507254	*CCL17*	*PSTPIP1*	*FHL1*
SEMA6A	*TUBA4A*	*EPOR*	*EEPD1*
PRX	*LTB*	*PON2*	*SIGLEC15*
PLEKHA5	*PYHIN1*	*PEAK1*	*B3GNT5*
LGMN	*SEL1L3*	*SCHIP1*	*CEBPE*
S100Z	*TNFRSF13C*	*TCF7*	*PLEKHG1*
LINC00114	*FCRL1*	*BAALC*	*DSG2*
MS4A1	*MPO*	*PSTPIP2*	*LIMS2*
GYLTL1B	*MYO7B*	*EBF4*	*LAX1*
ATP10A	*LOXHD1*	*SGSH*	*CAMK2D*
C1QTNF4	*C1orf228*	*CNN2*	*CD97*
ARHGEF17	*TNS1*	*PLK2*	*CRIP1*
ARHGEF4	*LRRC14B*	*MXRA7*	*USP32P1*
NCF2	*MARCKS*	*PRKAR2B*	*RGS10*
CD27	*CHN2*	*GPR160*	*TMEM156*
VWA2	*IGJ*	*SEMA3F*	*MAGEF1*
FSCN1	*PLIN2*	*FGFR1*	*GSN*
BMP2	*CXCR7*	*FYB*	*MYOCD*
LPCAT2	*HIST1H2BD*	*FGD2*	
KCNE3	*RCN3*	*NRXN3*	
TMEM154	*PROM1*	*AFAP1L2*	
NRP1	*PLVAP*	*GPRIN3*	
IGF2BP1	*DOK4*	*ICAM3*	
TMED6	*CD93*	*IGF2BP2*	
NRN1	*EPHA7*	*IL1RAP*	
DENND3	*PIM1*	*GAPT*	
PTP4A3	*LST1*	*METRNL*	
TMEM217	*SIRPA*	*BANK1*	

Supplemental Figures

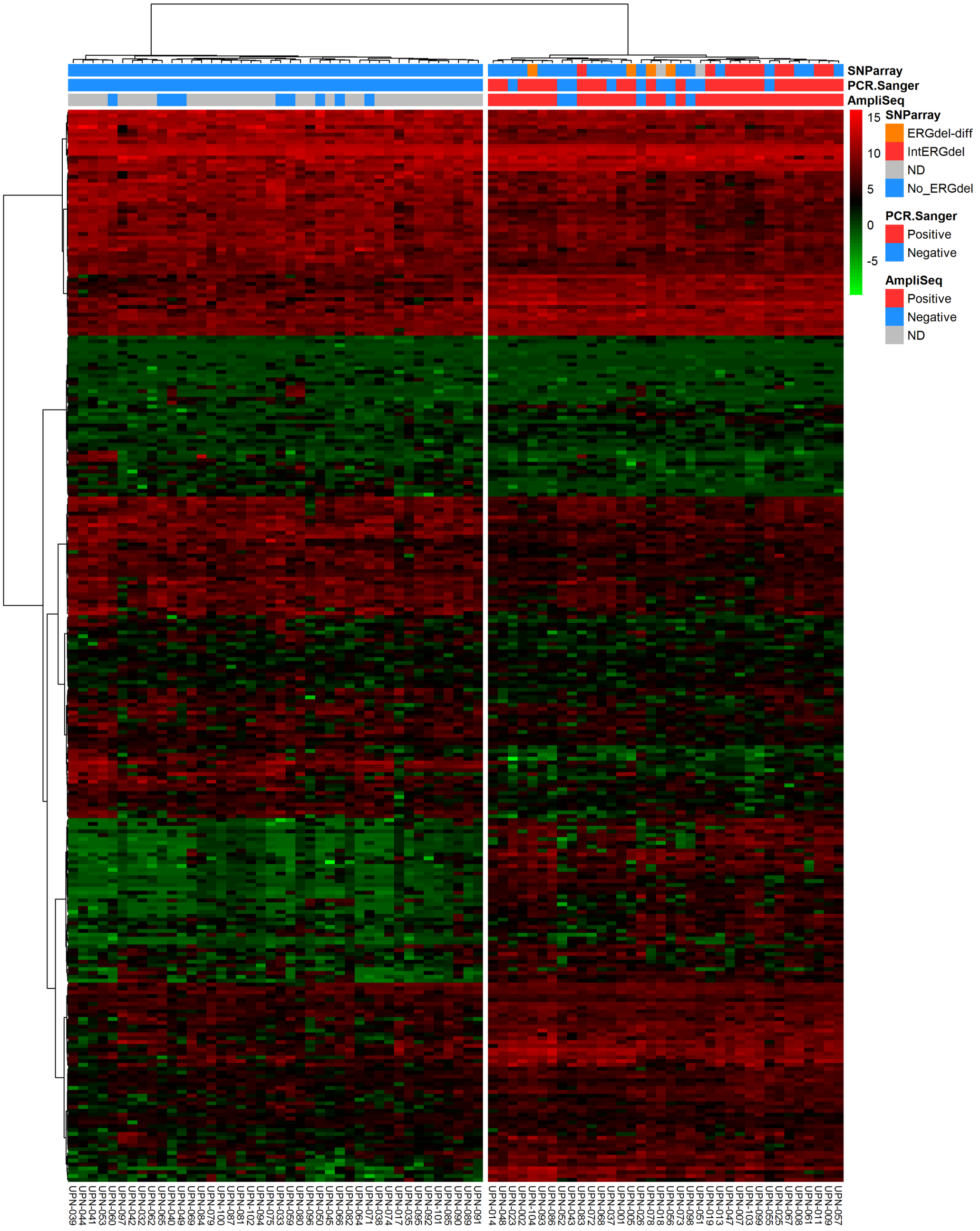
Supplemental Figure 1. Hierarchical clustering of 112 B-other ALL patients based on the expression of DUX4r-ALL signature genes analyzed by RNAseq

Forty-three patients in the left cluster were assigned to DUX4r-ALL. Patients are annotated with the results of the ERG gene deletion (ERGdel) screening by all 3 used methods. ND – not done.



Supplemental Figure 2. Hierarchical clustering of 78 B-other ALL patients based on the expression of DUX4r-ALL signature genes analyzed on microarray

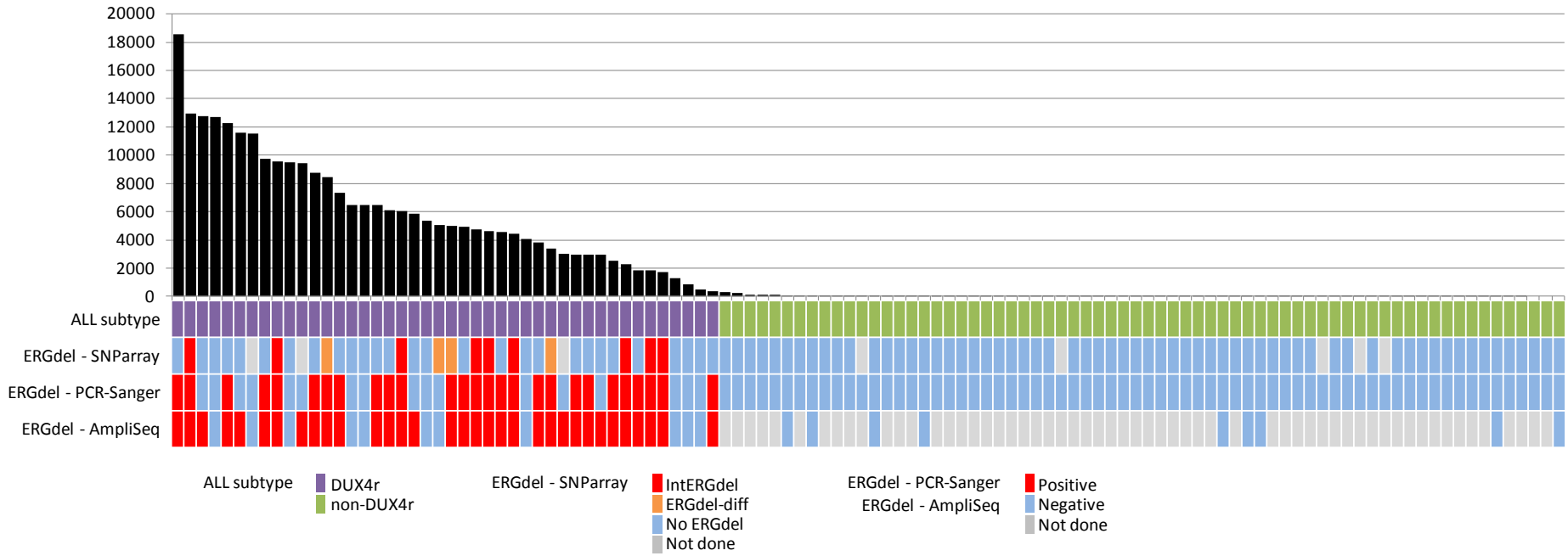
Thirty-six patients in the right cluster were assigned to DUX4r-ALL. Patients are annotated with the results of the ERG gene deletion (ERGdel) screening by all 3 used methods. ND – not done



Supplemental Figure 3. The DUX4 expression levels in 112 patients analyzed by RNAseq

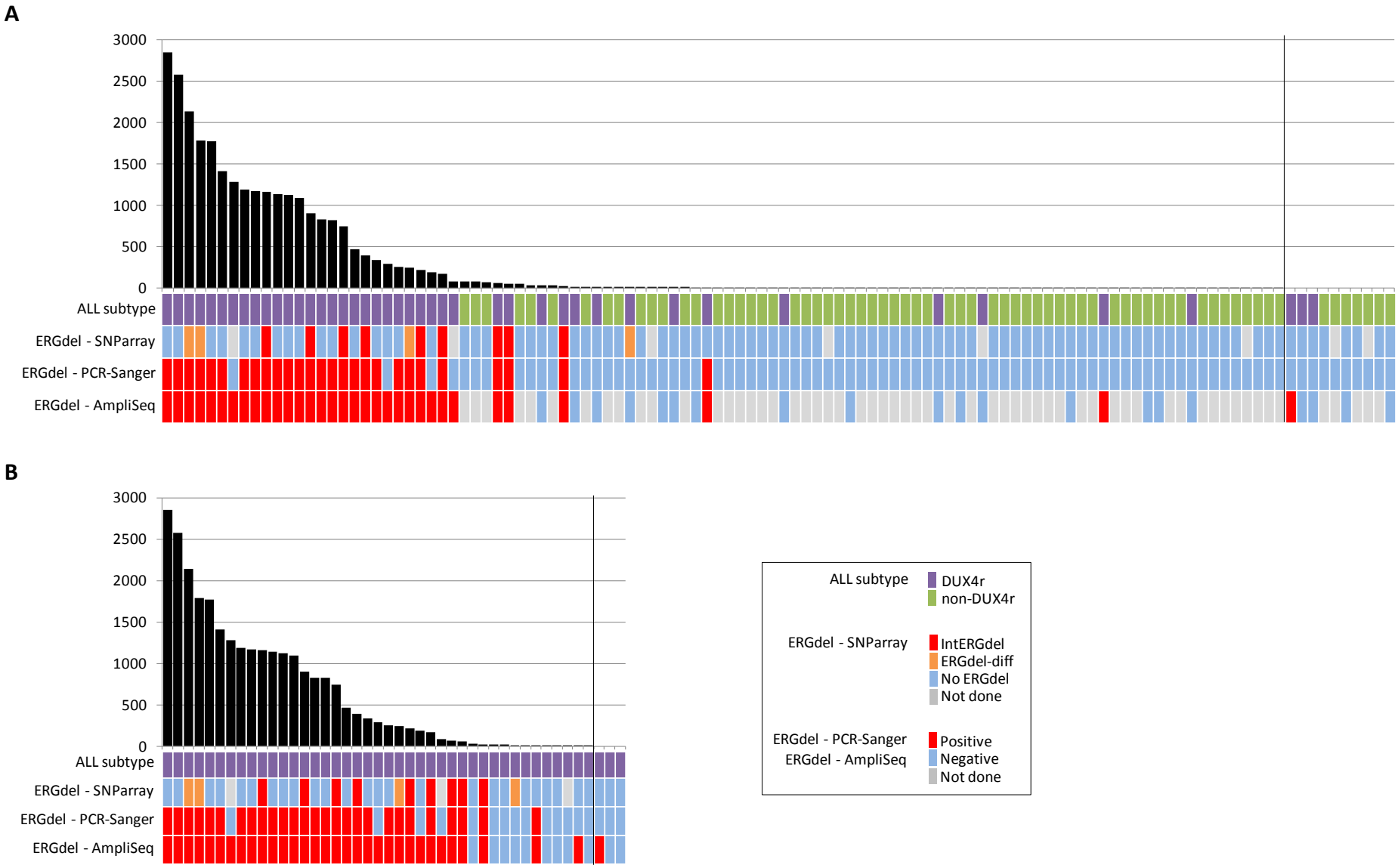
Patients are ranked according to the expression level (starting with the highest expression level on the left). All 44 *DUX4*r ALL patients have higher expression levels compared to 68 non-*DUX4*r ALL patients. The results of *ERG*del screening from SNP array, PCR + Sanger and AmpliSeq are shown for each patient.

y-axis: normalized read counts; x-axis: individual patients.



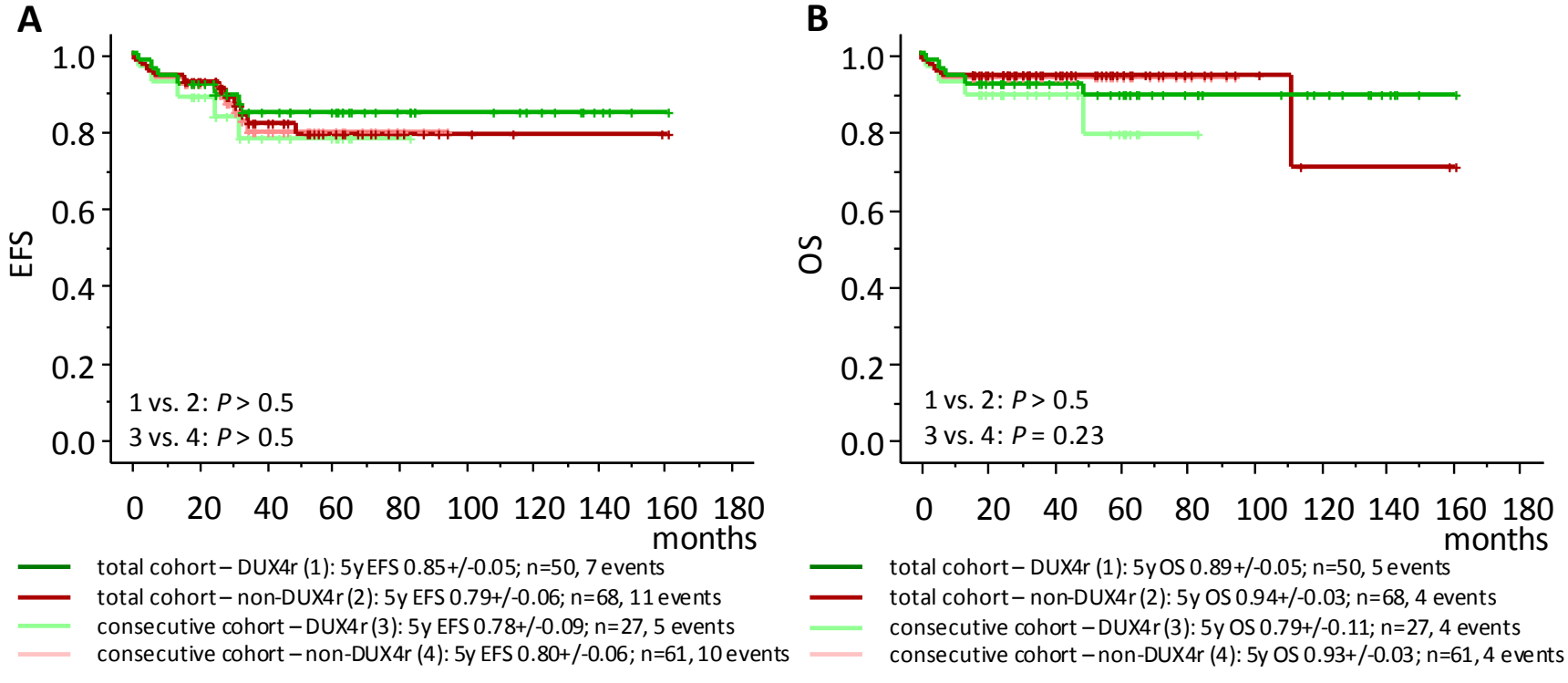
Supplemental Figure 4. The expression of ERGalt analyzed by RNAseq

(A-B) Graphs show the normalized numbers of reads containing exon junctions corresponding to *ERG*alt a and *ERG*alt b transcripts fom RNAseq. (A) Graph includes all 112 cases with RNAseq data, cases are ranked according to the numbers of junction reads in descending order; cases on the right of the vertical black line do not have any *ERG*alt junction reads and are grouped according to the ALL subtype. (B) Graph shows only 44 *DUX4*r cases with RNAseq data; cases are ranked according to the number of junction reads in descending order; cases on the right of the vertical black line do not have any *ERG*alt junction reads and are grouped according to the *ERG*del presence.



Supplemental Figure 5. Impact of DUX4r on survival within B-other ALL

5-year EFS (A) and OS (B) of all analyzed B-other patients stratified according to the DUX4r-ALL classification.



Supplemental Figure 6. Impact of ERGdel on survival within B-other ALL

5-year EFS (A, C, E) and OS (B, D, F) of all analyzed B-other patients stratified according to the ERGdel presence defined by SNP array (A, B), genomic PCR (C, D) and by any ERGdel positivity (SNP array, PCR, AmpliSeq) (E, F).