

Supplementary Figures and Tables

A convergent malignant phenotype in B-cell acute lymphoblastic leukemia involving the splicing factor SRRM1

Adria Closa^{1,2}, Marina Reixachs-Solé^{1,2}, Antonio C. Fuentes-Fayos^{3,4,5}, Katharina E. Hayer⁶, Juan Luis Melero^{1,2}, Fabienne R. S. Adriaanse⁷, Romy S. Bos⁷, Manuel Torres-Diz⁶, Stephen Hunger⁶, Kathryn G. Roberts⁸, Charles Mullighan⁸, Ronald W. Stam⁷, Andrei Thomas-Tikhonenko^{6,9}, Justo P. Castaño^{3,4,5,10}, Raúl M. Luque^{3,4,5,10}, Eduardo Eyras^{1,2,11,12,}*

¹The John Curtin School of Medical Research, Australian National University, Canberra, Australia.

²EMBL Australia Partner Laboratory Network at the Australian National University, Canberra, Australia.

³Maimonides Biomedical Research Institute of Cordoba (IMIBIC), Cordoba, Spain.

⁴University of Cordoba (UCO), Cordoba, Spain.

⁵Reina Sofía University Hospital, Cordoba, Spain.

⁶Children's Hospital of Philadelphia, Philadelphia, USA.

⁷Princess Máxima Center for Pediatric Oncology, Utrecht, Netherlands

⁸St. Jude Children's Research Hospital, Memphis, USA.

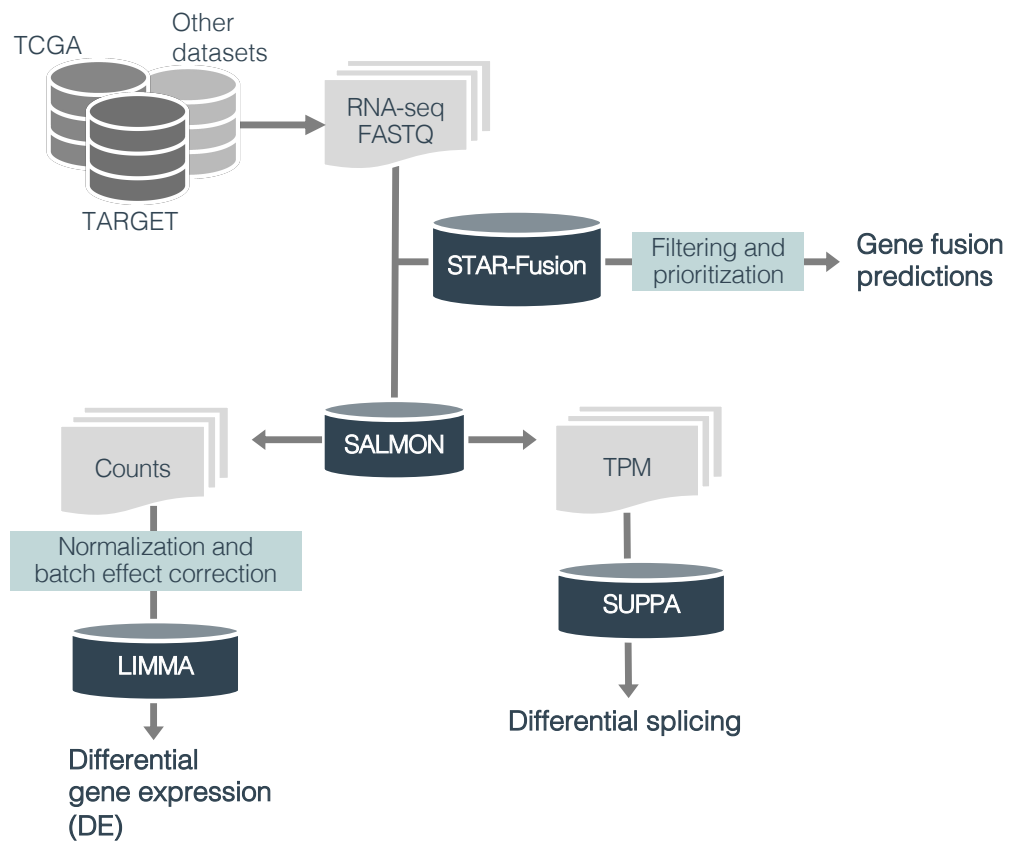
⁹University of Pennsylvania School of Medicine, Philadelphia, USA.

¹⁰Centro de Investigación Biomédica en Red de Fisiopatología de la Obesidad y Nutrición, (CIBERObn), Cordoba, Spain.

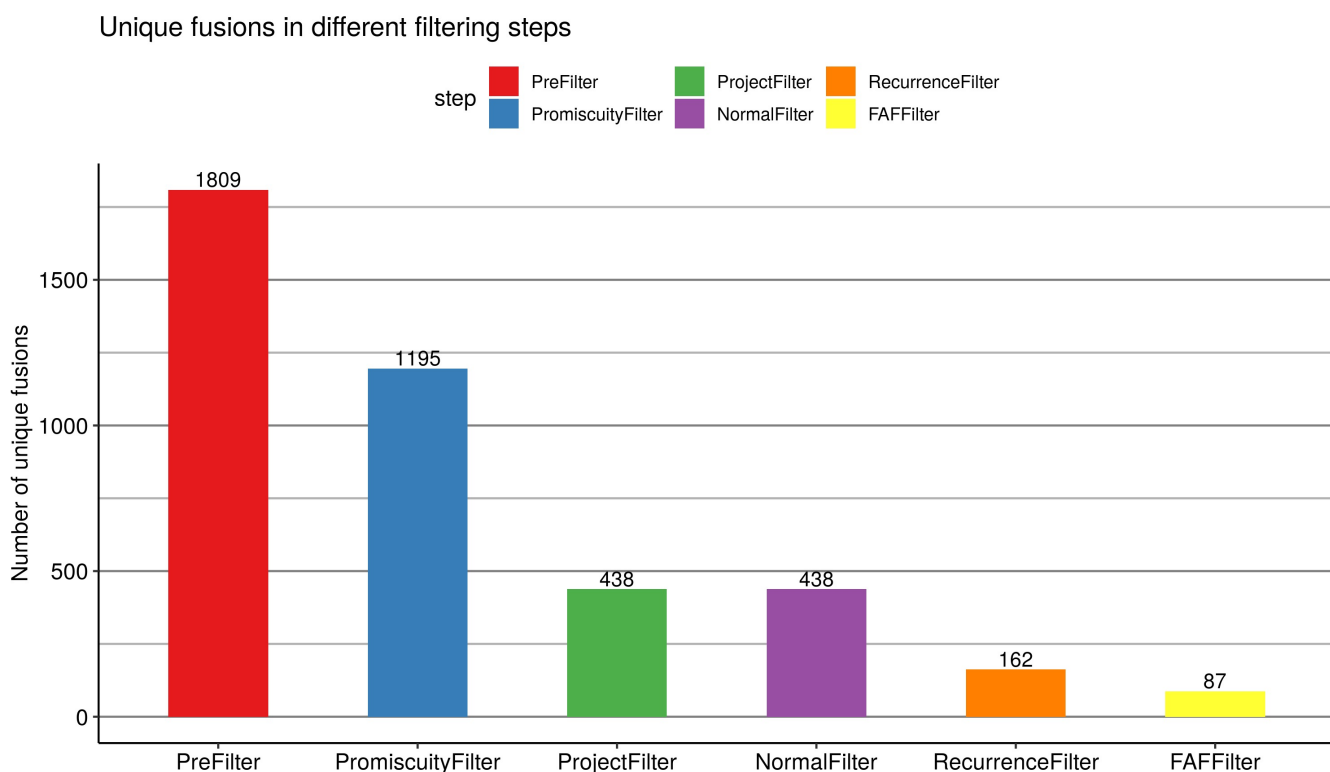
¹¹Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain.

¹²Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain.

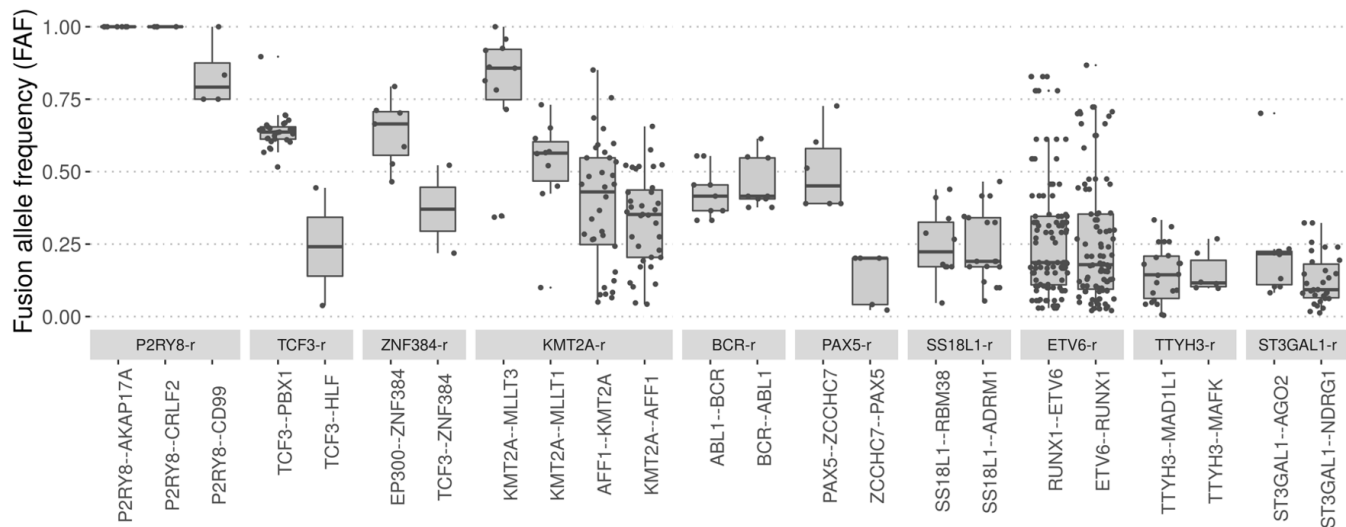
* correspondence to: eduardo.eyras@anu.edu.au



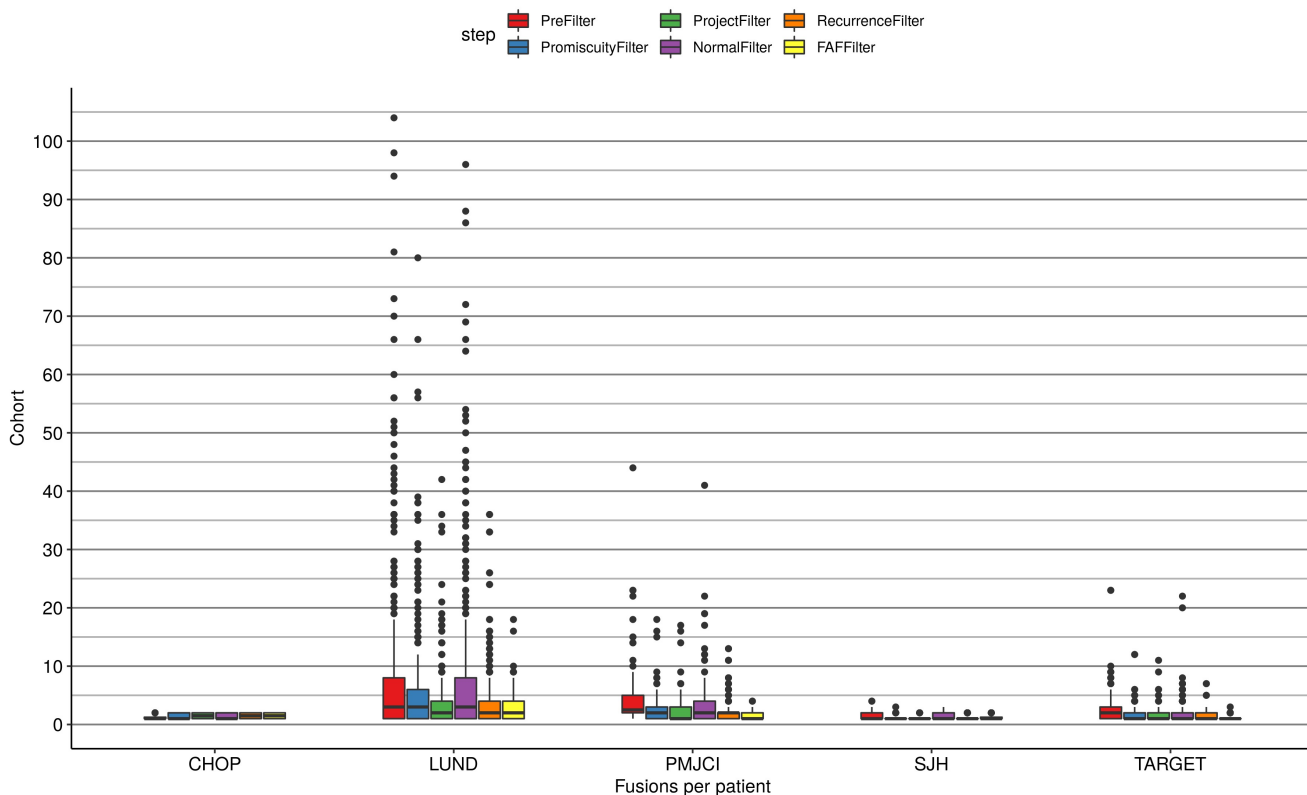
Supplementary Figure 1. RNA sequencing (RNA-seq) analysis pipeline. The figure depicts the analyses performed on the short-read RNA-seq samples from Data Table 1. The description of the methods used can be found in the Methods section.



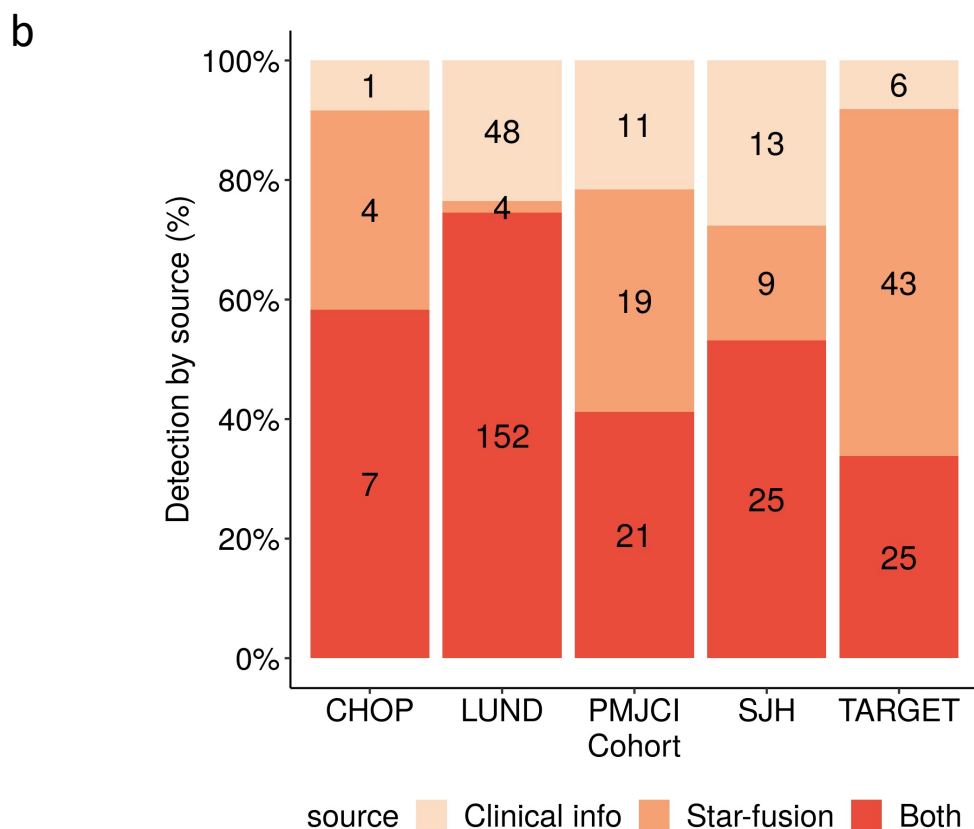
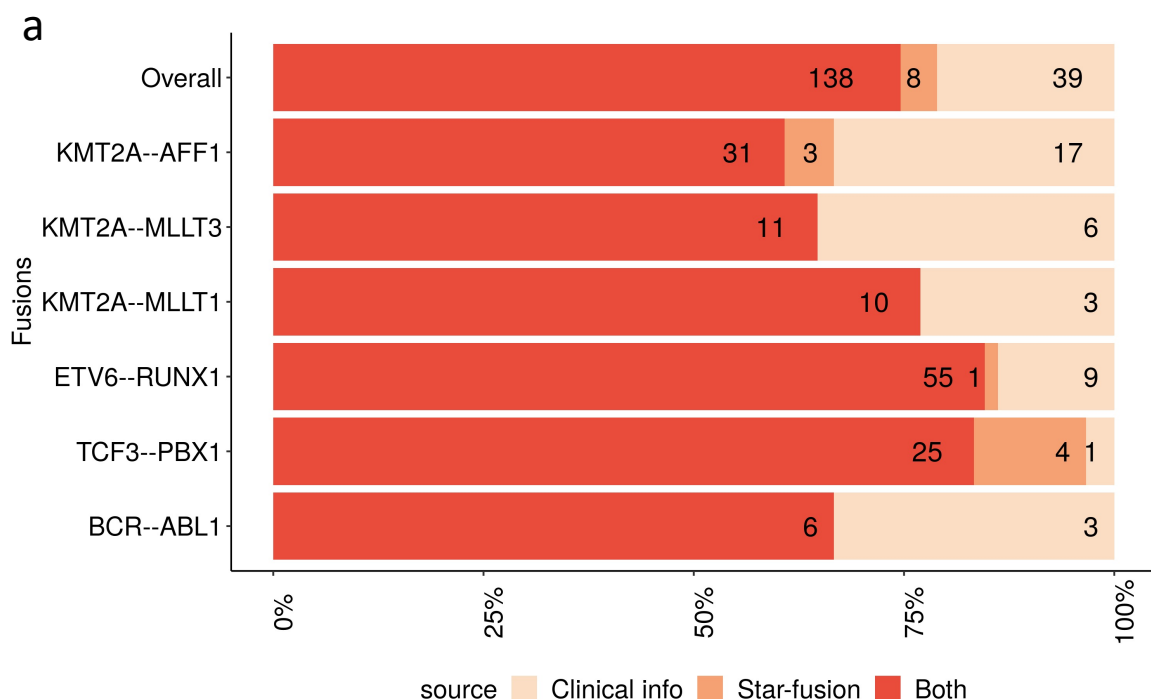
Supplementary Figure 2. Detection of gene-fusions. Bar plot showing the number of unique fusions identified after each filtering step was applied. We started with 1825 fusion candidates. On top of each bar, we indicate the number of fusions left after applying each filter consecutively, from left to right. PreFilter (red): fusions involving Ig genes, Hb genes, pseudogenes and paralogous genes were removed. PromiscuityFilter (blue): fusions involving genes with multiple partners in the same sample were removed, except if the fusions were previously observed in cancer. ProjectFilter (green): fusions that appeared in only one of the cohorts were removed. NormalFilter (purple): fusions appearing in normal samples were removed. RecurrenceFilter (orange): fusions occurring in fewer than 5 patients were removed, except for fusions involving genes that were observed before in other fusions or mutated in leukemia. FAFFilter (yellow): cases with low fusion allele frequency (FAF) were removed.



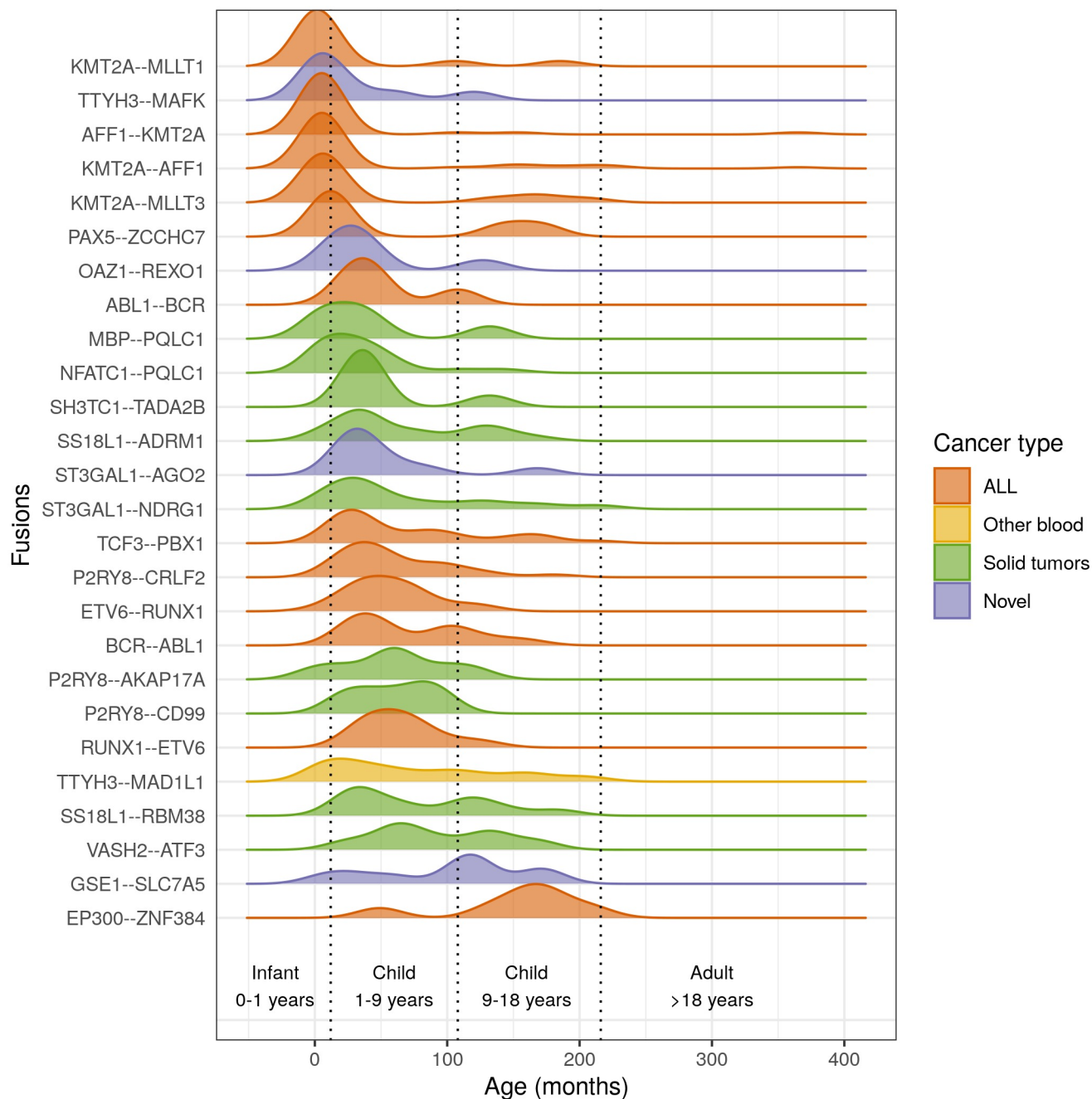
Supplementary Figure 3. Fusion allele frequency (FAF) values. For each of the fusion groups depicted in Figure 1b, we give the distribution of the FAF values (y axis) for the most common gene fusion pairs (x axis).



Supplementary Figure 4. Number of fusions per patient. Distribution of the number of fusions in each cohort after applying the filters described in Supp. Fig. 2. The fusions used in this work correspond to the yellow distributions, resulting from applying all filters.

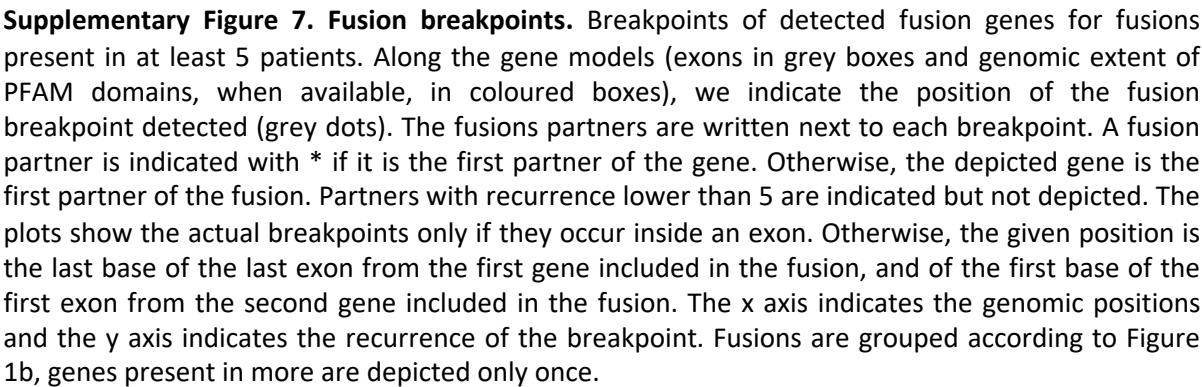
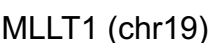
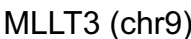
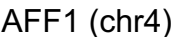


Supplementary Figure 5. Detection of annotated fusions measured by independent experimental methods. (a) For each fusion, we show the proportion of patients in which 1) the fusion was detected from RNA-seq and was annotated in the clinical information based on independent experimental methods (Both), 2) the fusion was only detected from RNA-seq (Star-fusion), and 3) the fusion was annotated in the clinical information but not detected in RNA-seq (Clinical info). **(b)** Proportion of fusions in each of the B-ALL cohorts separated as in (a).

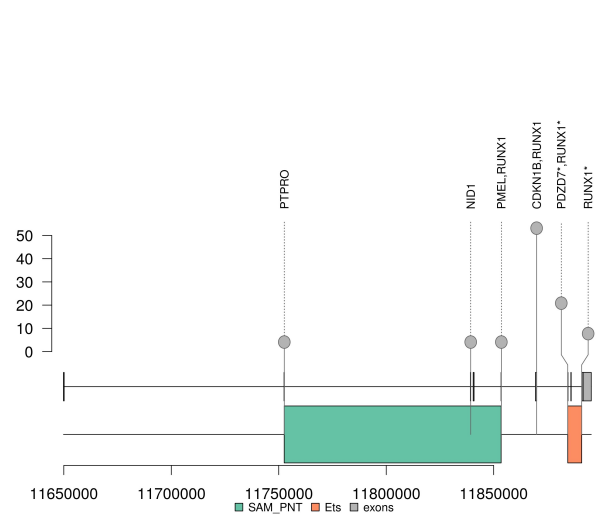


Supplementary Figure 6. Fusion range distribution by age. Age distribution of the patients (x axis) by fusion (y axis), colored according to whether the fusion has been previously described in ALL, other blood cancers (Other blood), solid tumors, or has not been previously described in any cancer type (Novel).

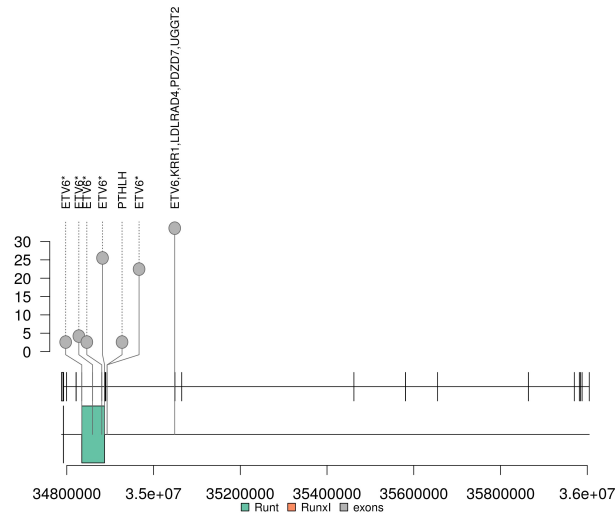
KMT2A (chr11)



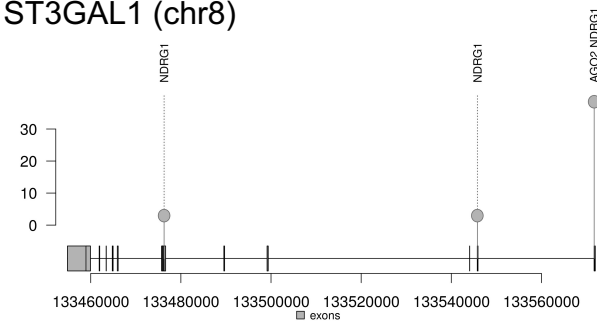
ETV6-r
ETV6 (chr12)



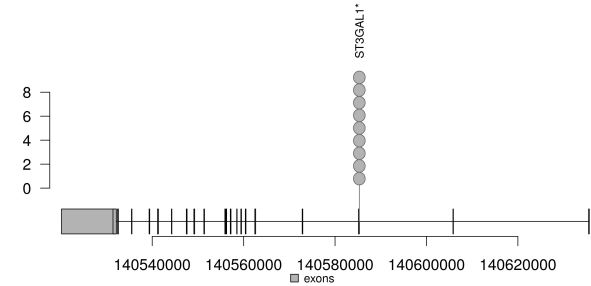
RUNX1 (chr21)



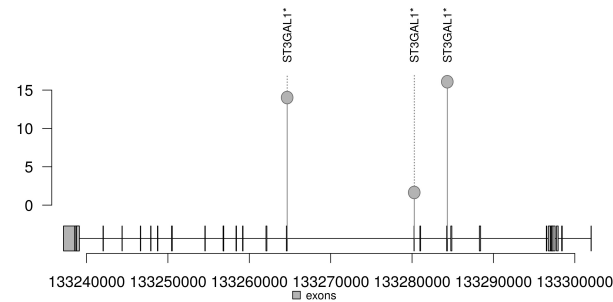
ST3GAL1-r
ST3GAL1 (chr8)



NDRG1 (chr8)

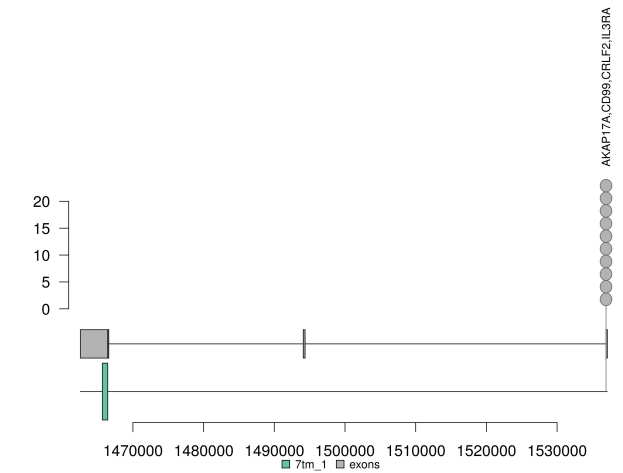


AGO2 (chr8)

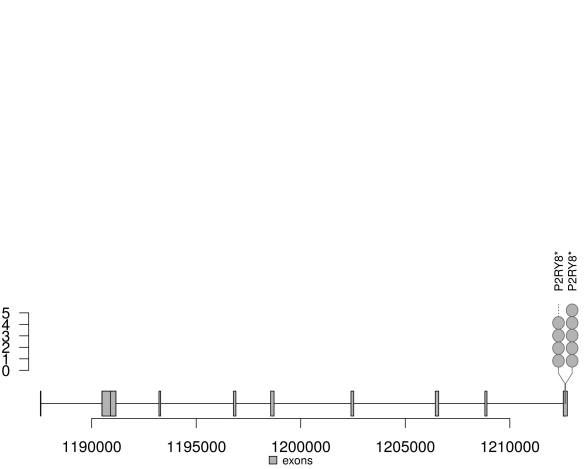


(Supplementary Figure 7 cont.)

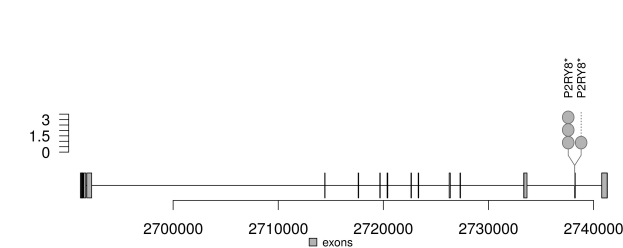
P2RY8-r
P2RY8 (chrX)



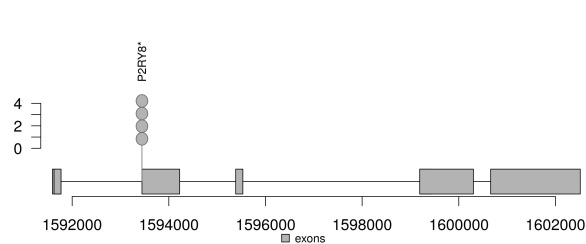
CRLF2 (chrX)



CD99 (chrX)

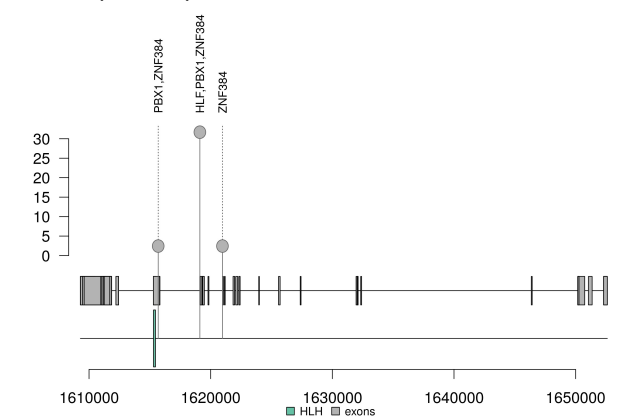


AKAP17A (chrX)

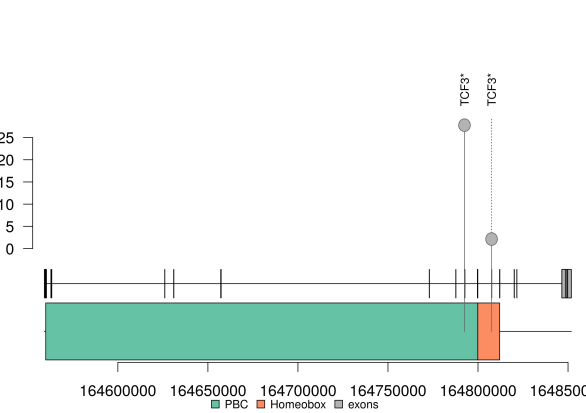


TCF3-r

TCF3 (chr19)

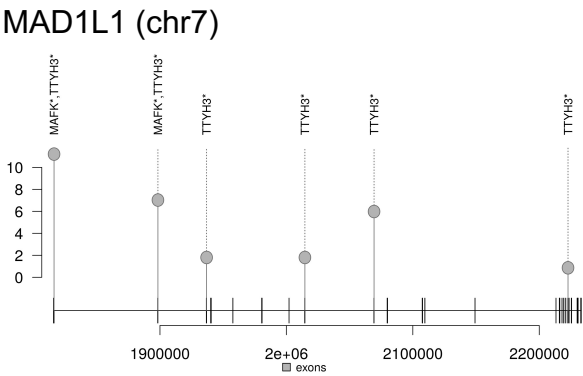
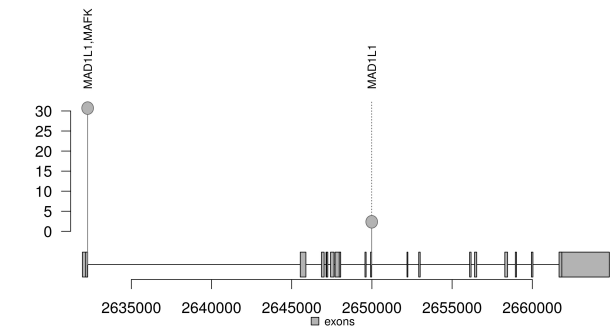


PBX1 (chr1)

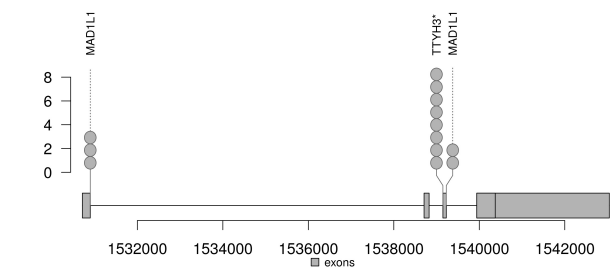


(Supplementary Figure 7 cont.)

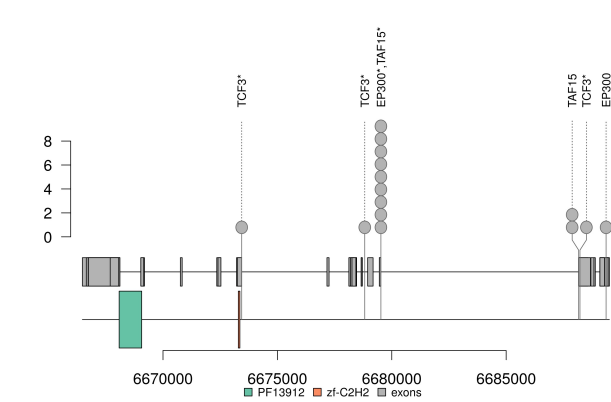
TTYH3-r TTYH3 (chr7)



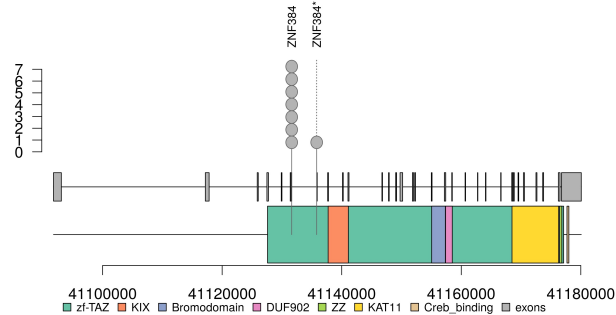
MAFK (chrX)



ZNF384-r ZNF384 (chr12)



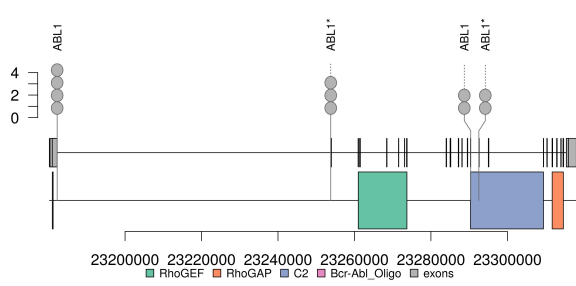
EP300 (chr22)



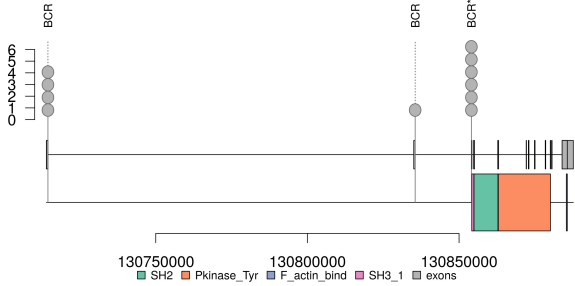
(Supplementary Figure 7 cont.)

BCR-r

BCR (chr22)

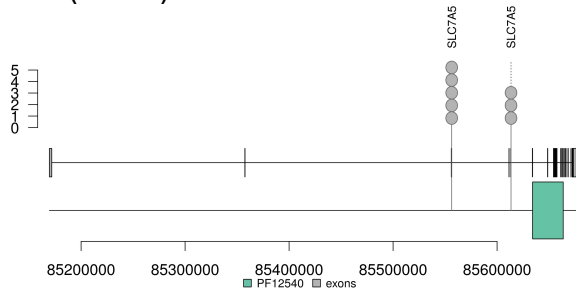


ABL1(chr9)

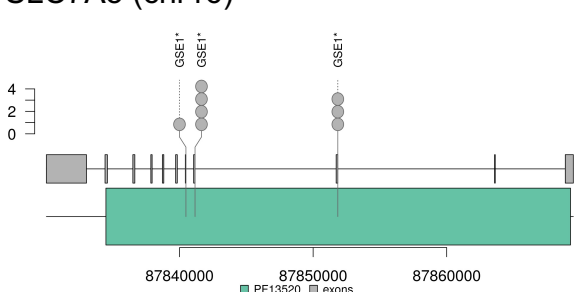


GSE1-r

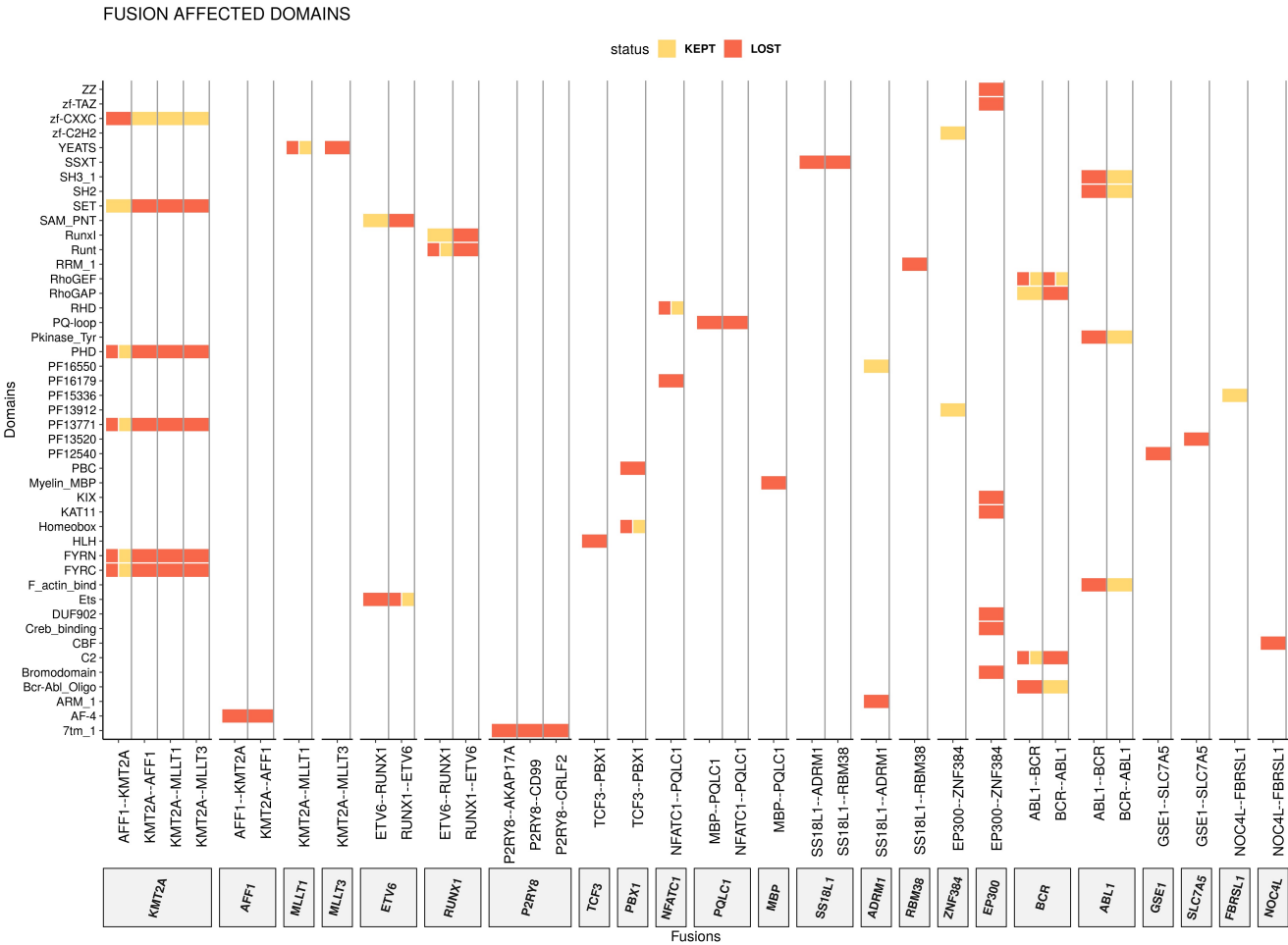
GSE1 (chr16)



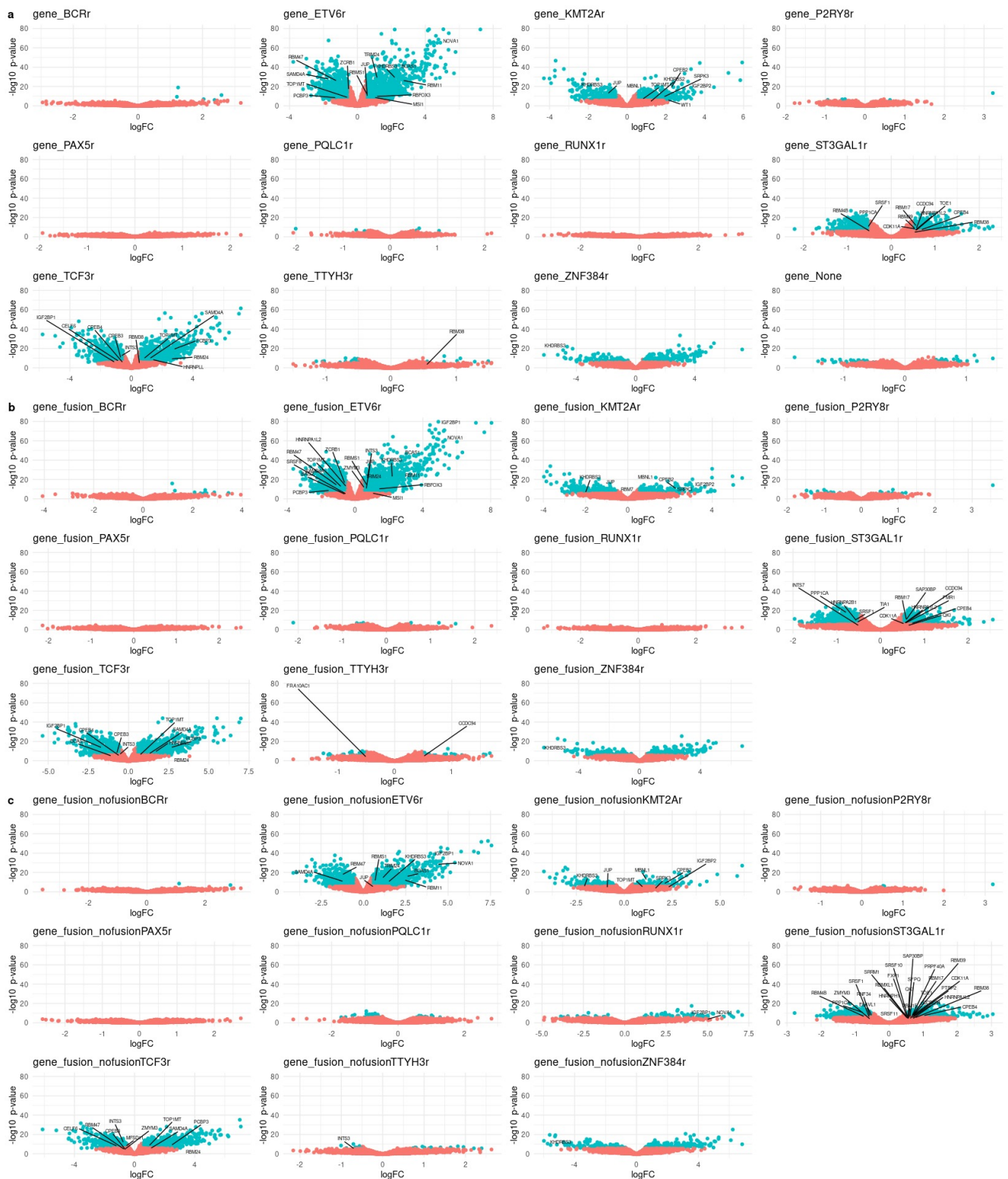
SLC7A5 (chr16)



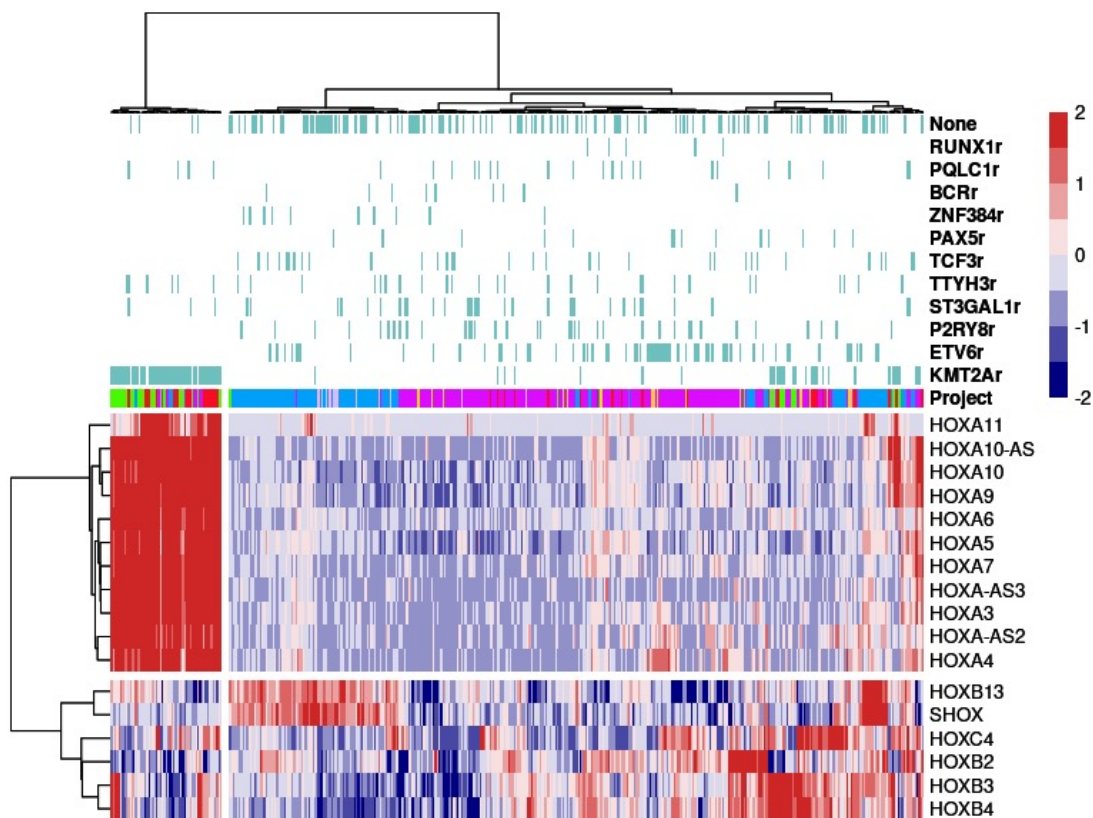
(Supplementary Figure 7 cont.)



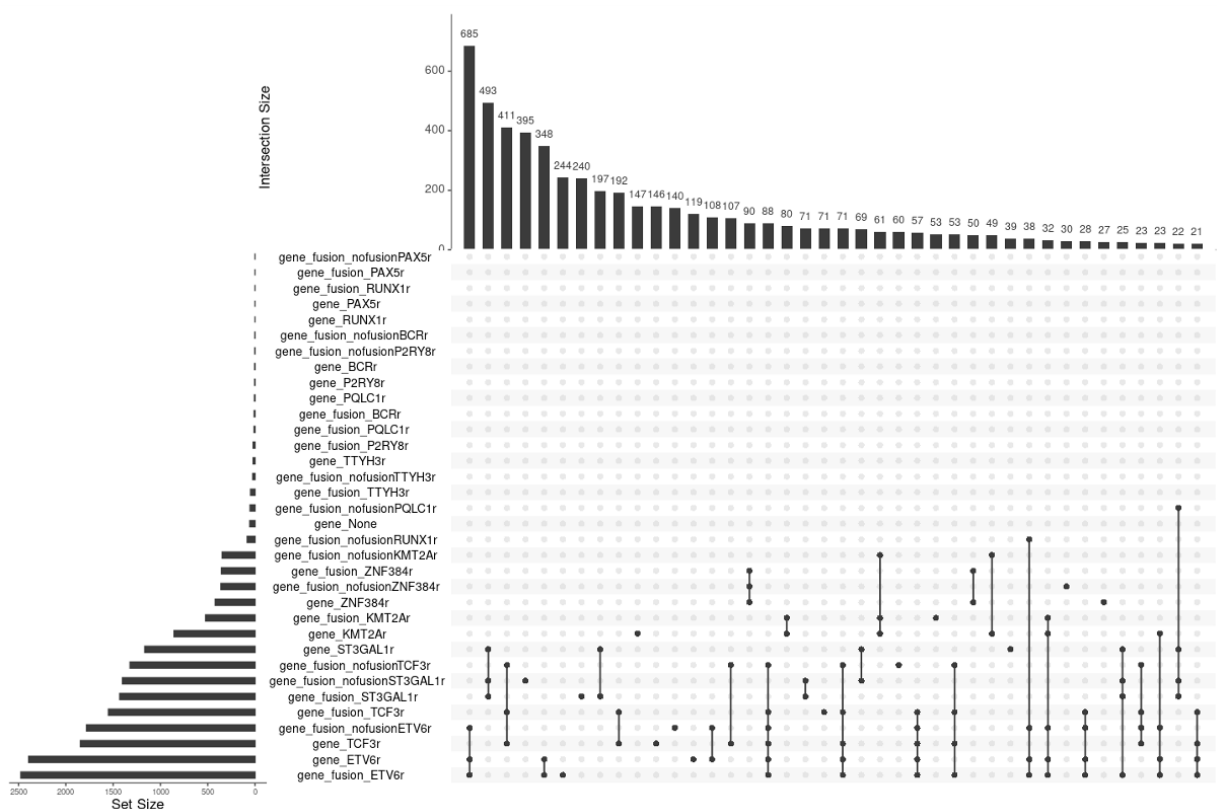
Supplementary Figure 8. Domains kept and lost through gene fusions. For each fusion (x axis), we indicate the domain (y axis) that is kept (yellow) or lost (red) in the fusion gene indicated in the box below the fusions.



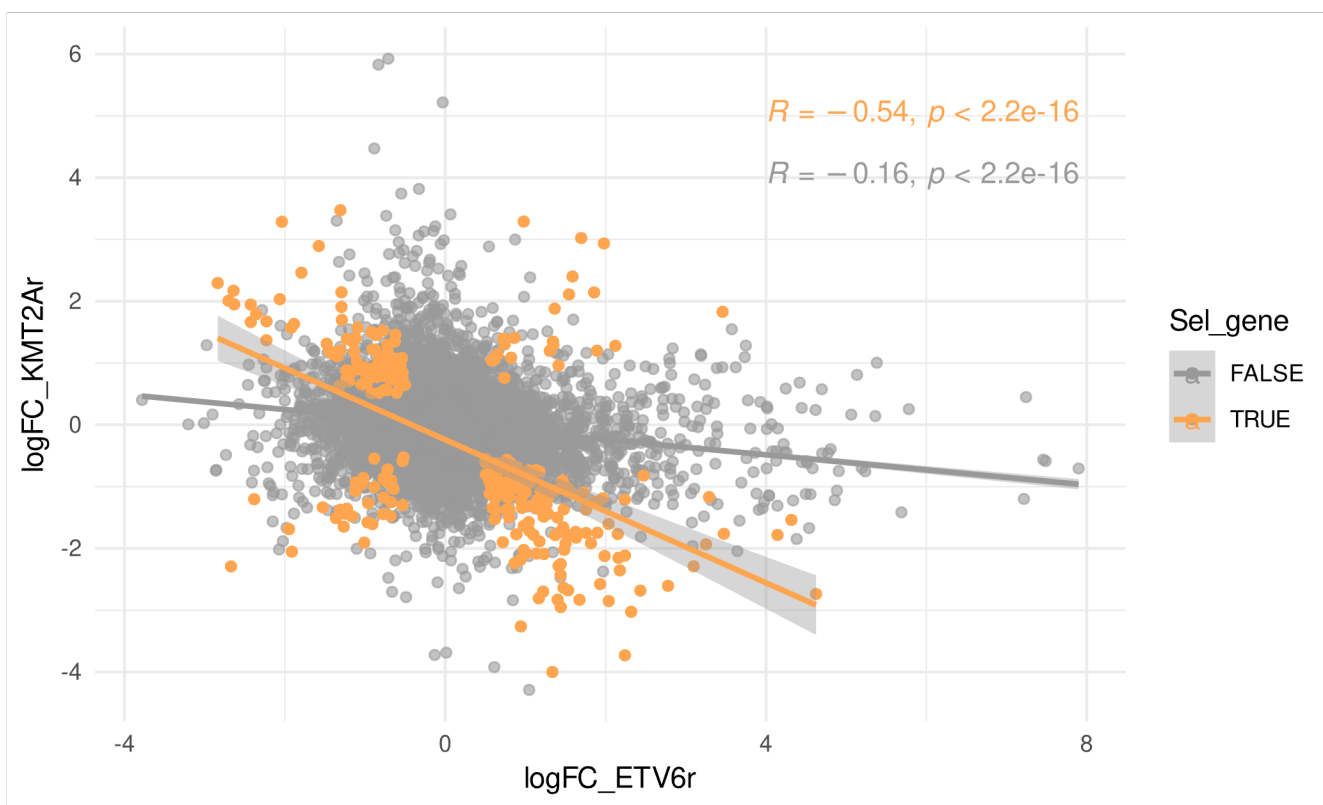
Supplementary Figure 9. Systematic differential expression analysis between fusion groups. For each fusion group (Supp. Table 2), we show the volcano plots for the differential expression analysis in three comparisons: **(a)** the fusion group vs. all leukemia patients (12 comparisons), **(b)** each fusion group vs. patients with some other fusion (11 comparisons), and **(c)** the fusion group vs. leukemia patients without fusion (11 comparisons). Patients that fall in more than one fusion group, because they have multiple fusions classified in different fusion groups, were removed from the comparisons. Volcano plots show on the x axis the log2 fold-change and on the y axis the $-\log_{10}(\text{corrected p-value})$.



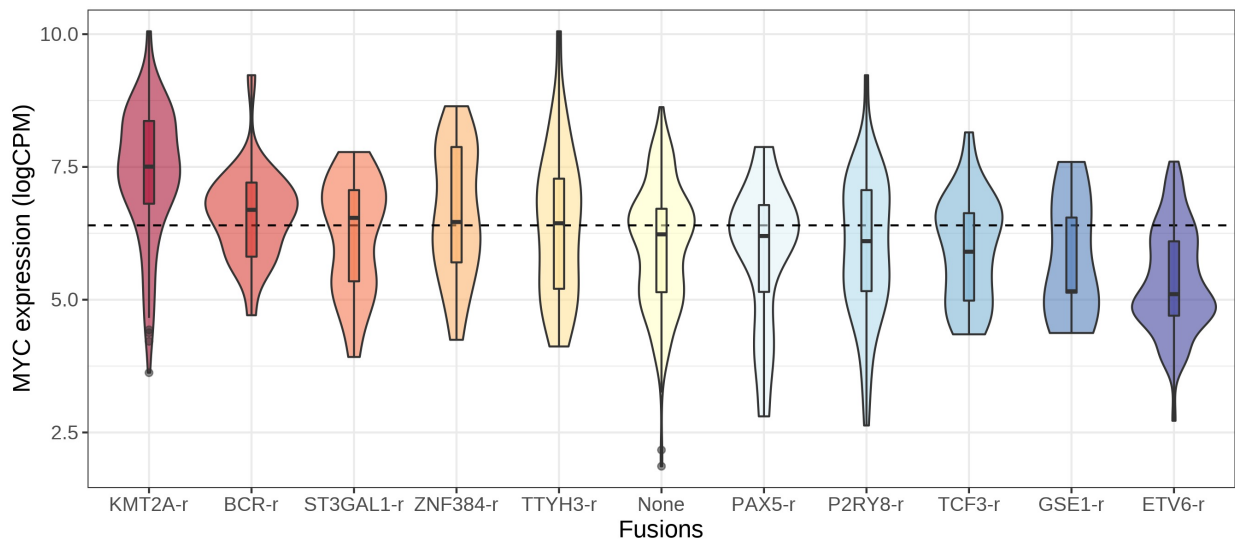
Supplementary Figure 10. Expression of HOX genes. We cluster all patients according to the expression of the HOX genes. Expression is given as the z-score of the log2(CPM) (CPM: counts per million). For each patient, we indicate above their cohort (Project) and fusion group. We indicate with “None” those patients without any recorded fusion.



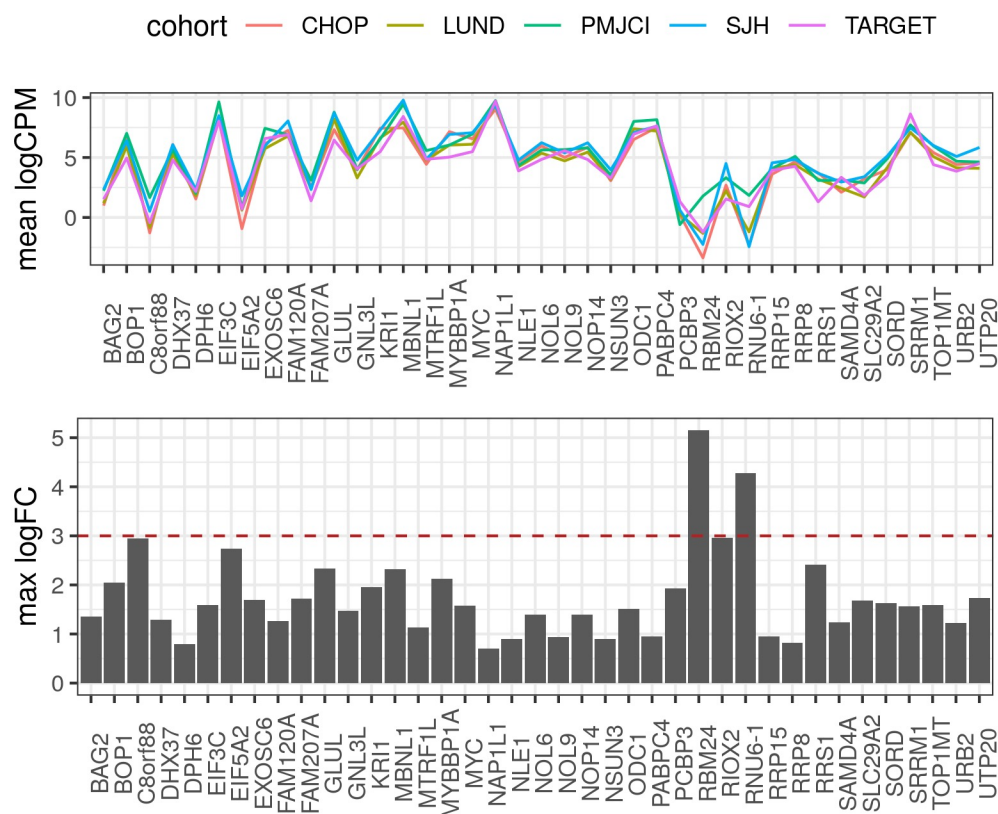
Supplementary Figure 11. Overlaps between differentially expressed genes in each fusion group. Intersection of genes differentially expressed from each comparison in Supp. Fig. 9. For each comparison on the left, the horizontal bar plot indicates the number of differentially expressed genes, and the vertical bar plot indicates the intersection of genes in two or more of these comparisons (indicated by bullet points). The intersections did not consider the direction of change.



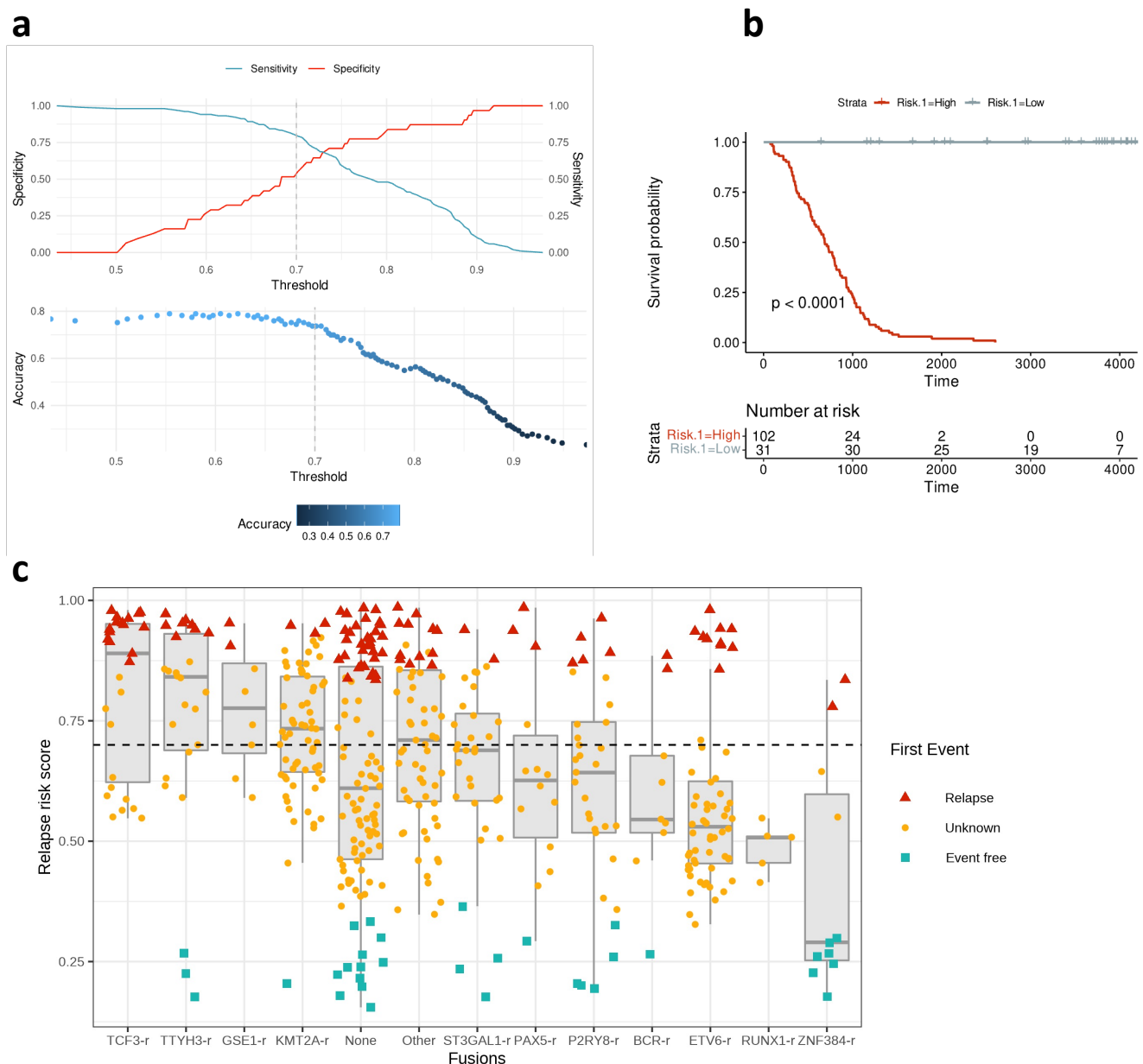
Supplementary Figure 12. Overlap of genes that are differential expressed between *KMT2A-r* or *ETV6-r* and the other groups of leukemia patients. The scatterplot shows the log2 fold-change (logFC) from comparing *ETV6-r* with the rest of leukemia patients (x axis), and from comparing *KMT2A-r* with the rest of leukemia patients (y axis). In orange we show the genes that were significant in both comparisons, in grey all the genes that showed no significance in any of the comparisons. R indicates the Pearson correlation value.



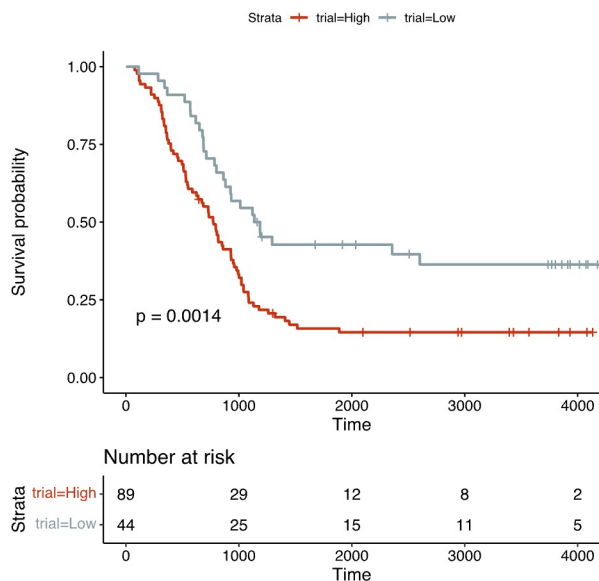
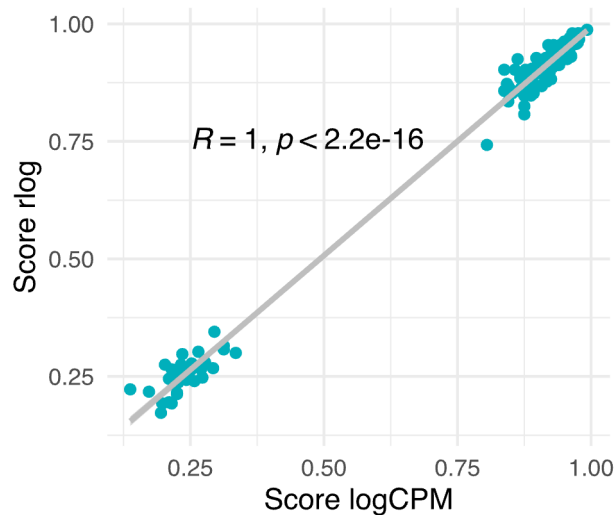
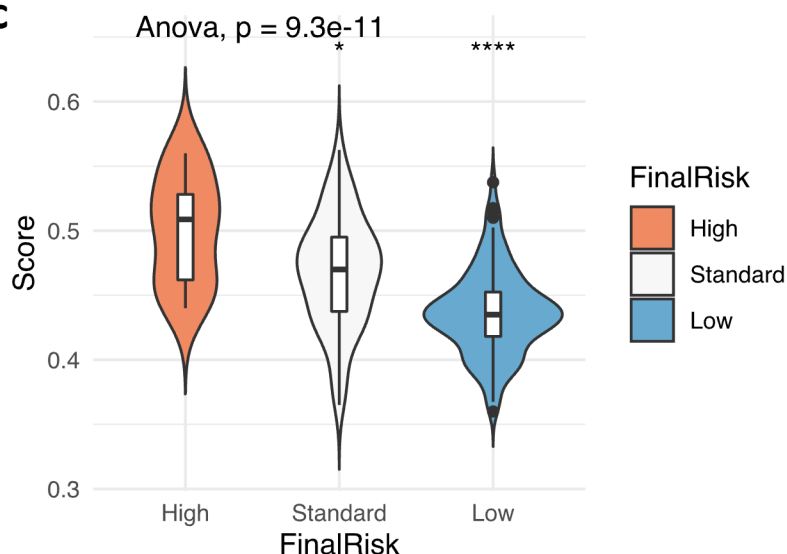
Supplementary Figure 14. Distribution of MYC expression values (y axis) for each fusion group compared to the mean expression in normal Feta-liver B-cells (horizontal dashed line). Expression is represented in $\log_2(\text{CPM})$, where CPM: counts per million.



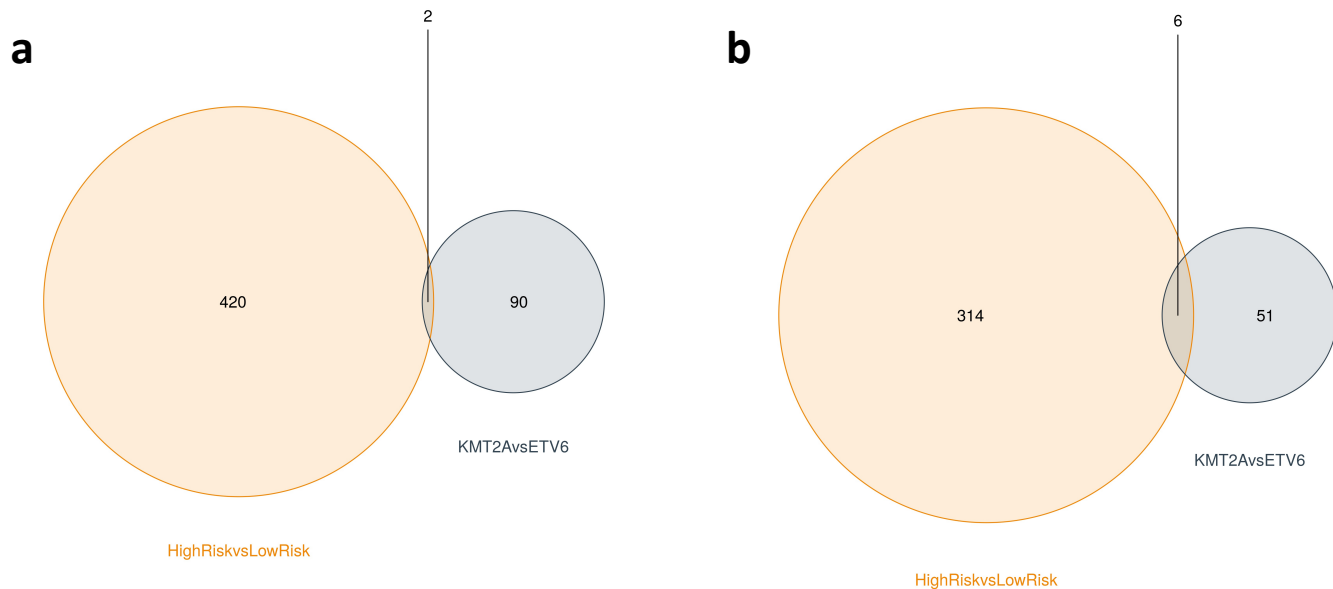
Supplementary Figure 15. Inter cohort gene variation from 39 gene candidates. Top panel show the mean logCPM expression for every gene by cohort. Bottom panel show the maximum logFC between the mean logCPM expression by cohort.



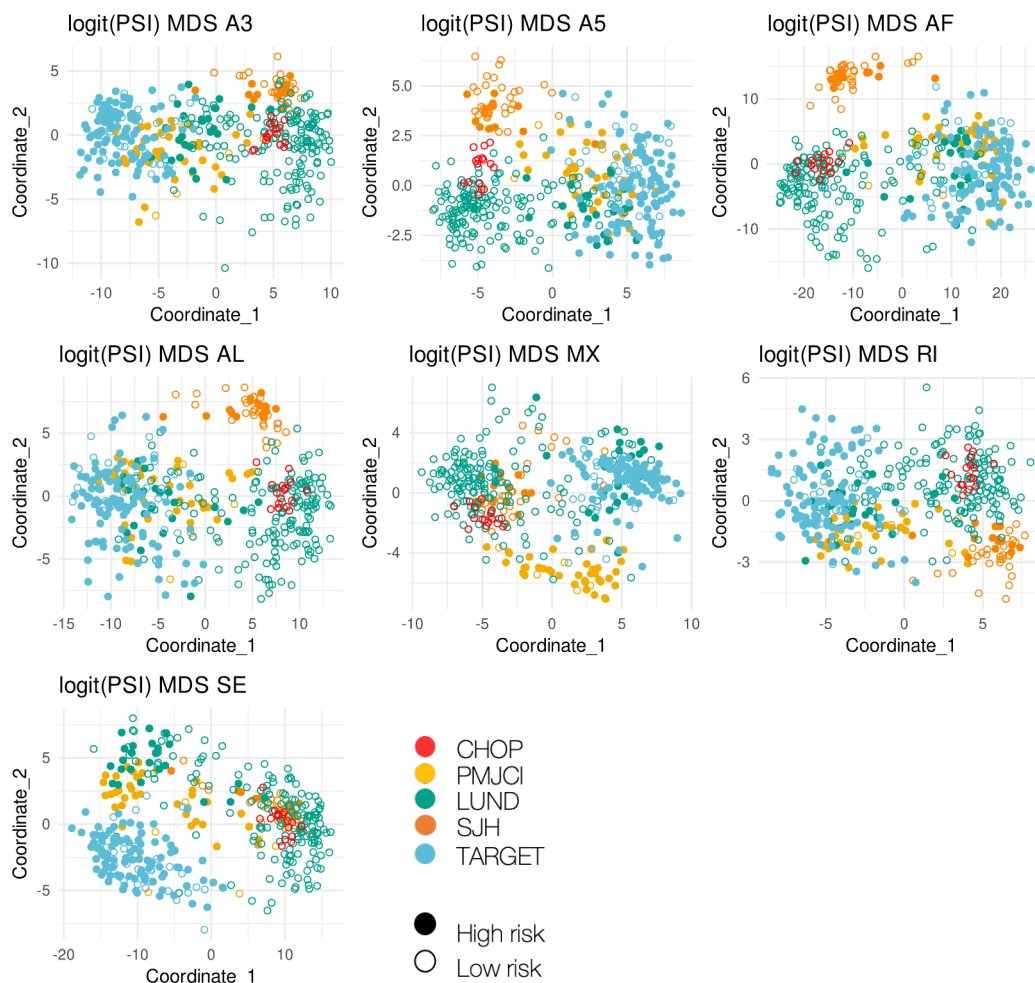
Supplementary Figure 16. Gene expression signature associated to high risk. (a) Specificity, Sensitivity, and Accuracy of the classification into low and high risk (y axis) as a function of the model score (x axis). We selected the threshold at score = 0.7, which shows a good balance between Specificity, Sensitivity, and Accuracy. Sensitivity is here calculated as the proportion of high-risk cases that are correctly predicted. Specificity is here calculated as the proportion of low-risk cases that are correctly predicted as low-risk. Accuracy is the proportion of correct predictions (high or low risk cases) over the total number of cases. **(b)** Kaplan-Meier curve of the patients separated by the model, prior to performing leave-one-out benchmarking. **(c)** Box plots with the distribution of the K-score values for each of the fusion groups tested. We indicate the used threshold of 0.7 with a dashed line. Cases with relapse are indicated as triangles, cases with no relapse as squares, and cases without follow up annotation are indicated as circles.

a**b****c**

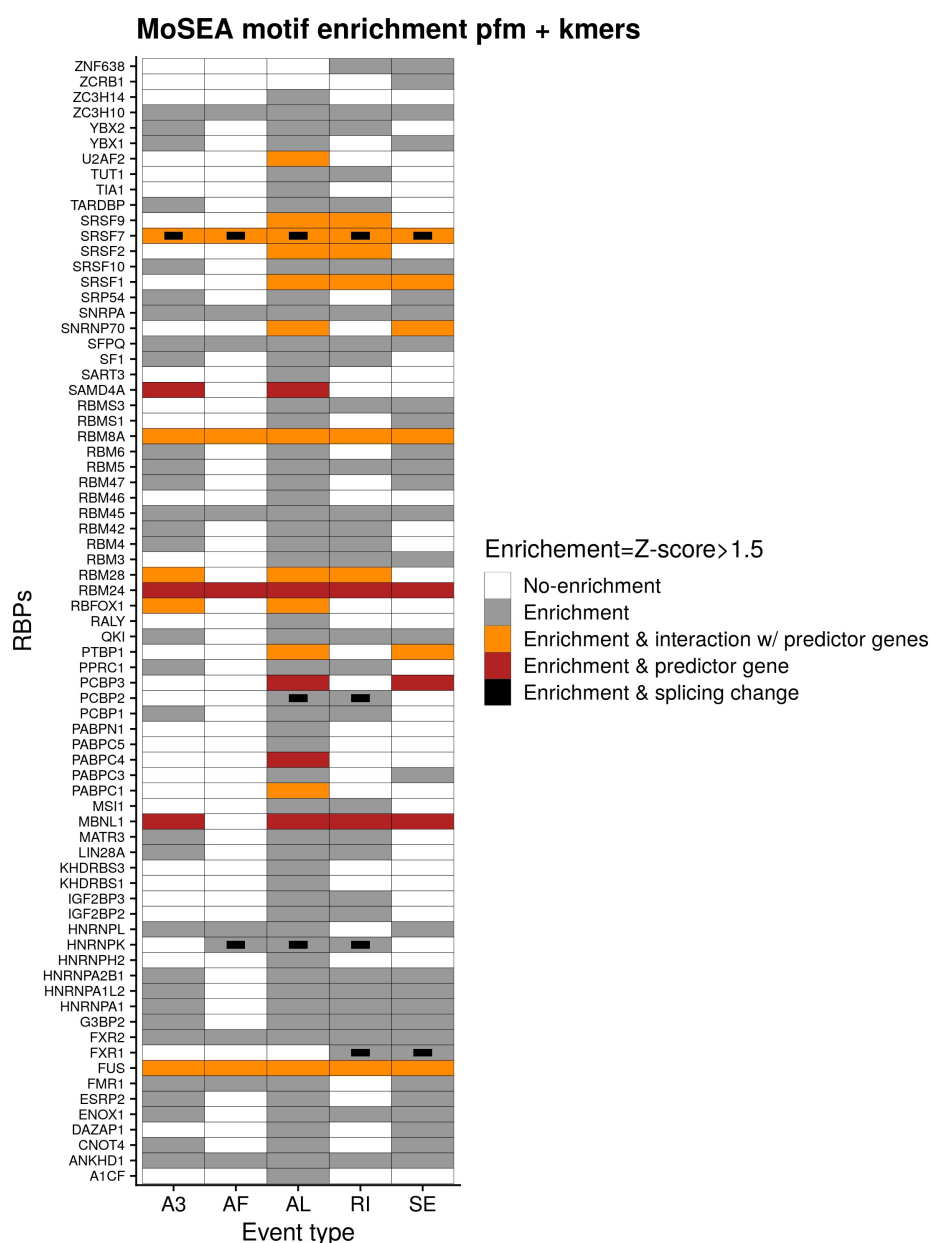
Supplementary Figure 17. Signature validation on an independent cohort. a) Kaplan-Meier plot of the patients separated as high risk (red) (risk score ≥ 0.7) or low risk (grey) (risk score < 0.7) in a leave-one-out test with the rlog model. **b)** Pearson correlation between target scores using logCPM model and rlog model. **c)** Violin plots with patients scores from an independent cohort using rlog model and separated by the risk groups described on the paper. One-way Anova test for global mean differences. T-test to compare means by group regarding High-risk group (* indicates $p \leq 0.05$, **** indicates $p \leq 0.00001$).



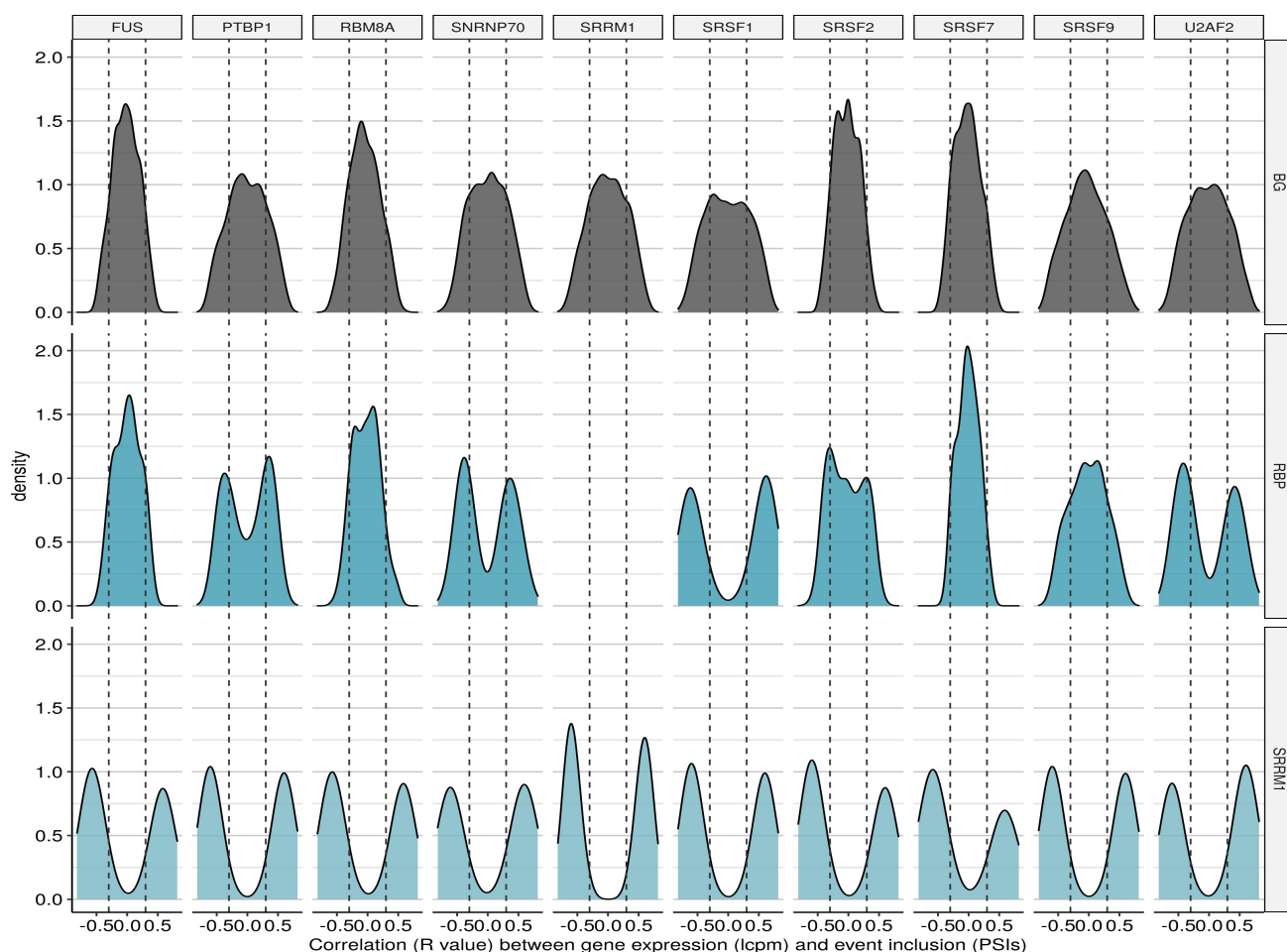
Supplementary Figure 18. Venn Diagram comparison high risk vs low risk and KMT2A-r vs ETV6-r. (a) Venn diagram overlapping differential included events comparing the High-risk vs low-risk patients from the predictor and comparing patients with KMT2A-r vs ETV6-r. **(b)** Venn diagram overlapping the genes affected by differential included events from the previous comparison on (a).



Supplementary Figure 19. Splicing events associated with high risk. The plot shows the multidimensional scale (MDS) analysis of the risk-associated events. Each dot represents a patient. The MDS was performed with the set of events associated with high risk, separated by the event type. Every color indicates the cohort of origin of the patient. A full dot is a patient predicted as high risk, whereas an empty dot is a patient predicted as low risk.

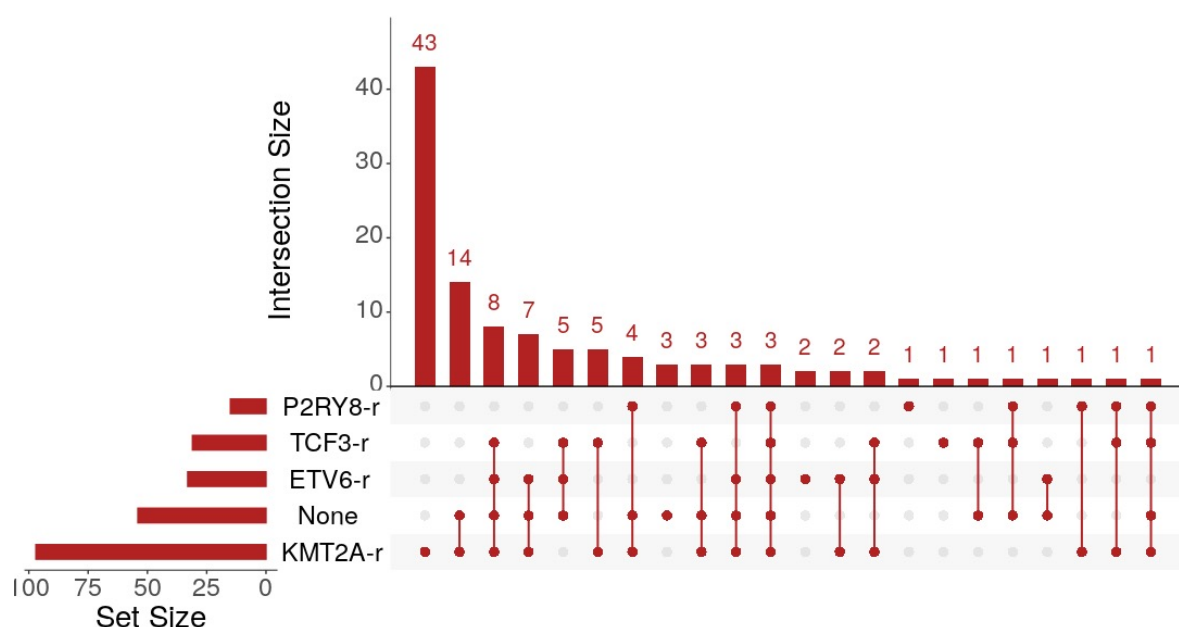


Supplementary Figure 20. Motif enrichment. Heatmap with the RBPs with enriched motifs in the events associated with high risk from the predictor. Motif enrichment was performed using MoSEA with k-mers and position frequency matrices (PFMs). MoSEA calculates a z-score to determine the association of the motif with the events that change inclusion between high and low risk, relative to the events that do not undergo any change. We indicate in white no enrichment, in grey when there is an enrichment, in orange when there is enrichment and the RBP interacts (via PPIs) with genes included in the predictor, and in red when there is motif enrichment and the RBP is part of the predictor. The black square inside indicates that there was a significant splicing change in the RBP gene.

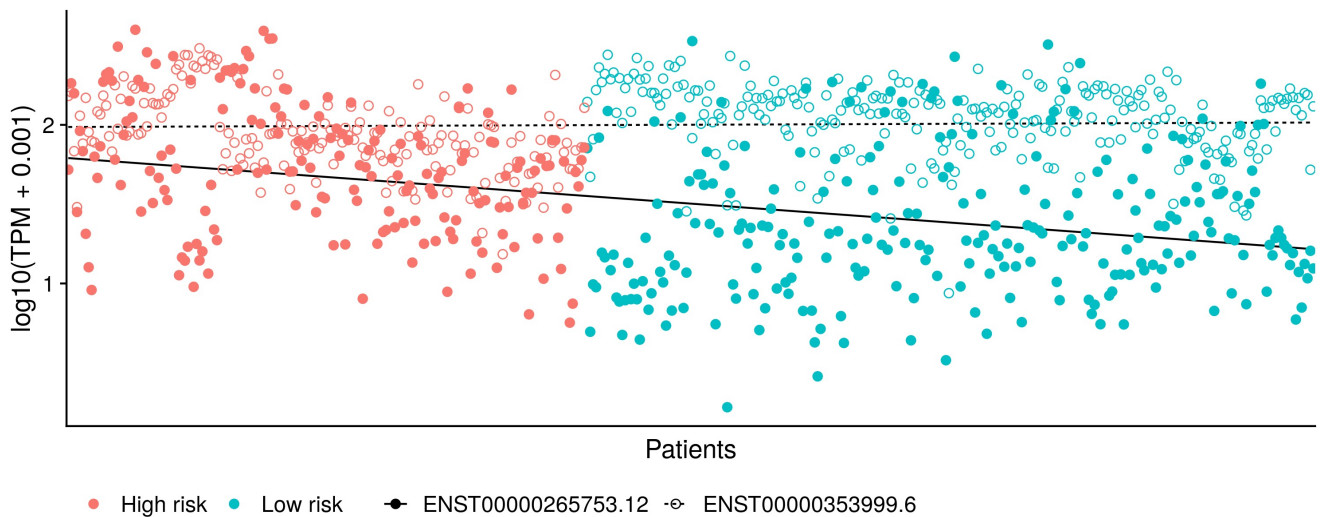


Supplementary Figure 21. Correlation plot using RNA processing events with RBPs and SRRM1.

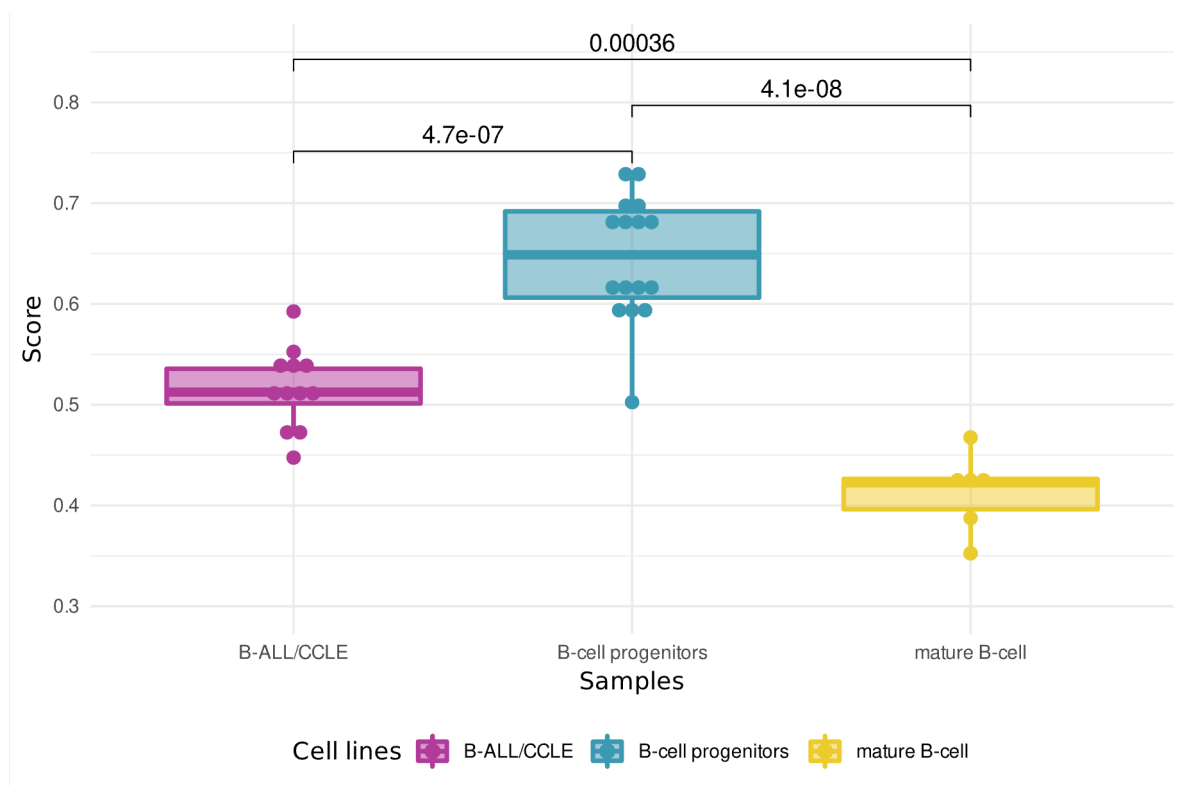
Distributions of the correlation values between splicing events and the expression of SRRM1 and the RBPs that interact with it along all the samples. The top row shows the distribution for background events randomly selected. The middle row show the events with a significant change between high and low risk and that also have a motif for the corresponding RBP. The bottom row show the same events from the middle row, but the distributions correspond to the correlation with SRRM1 expression.



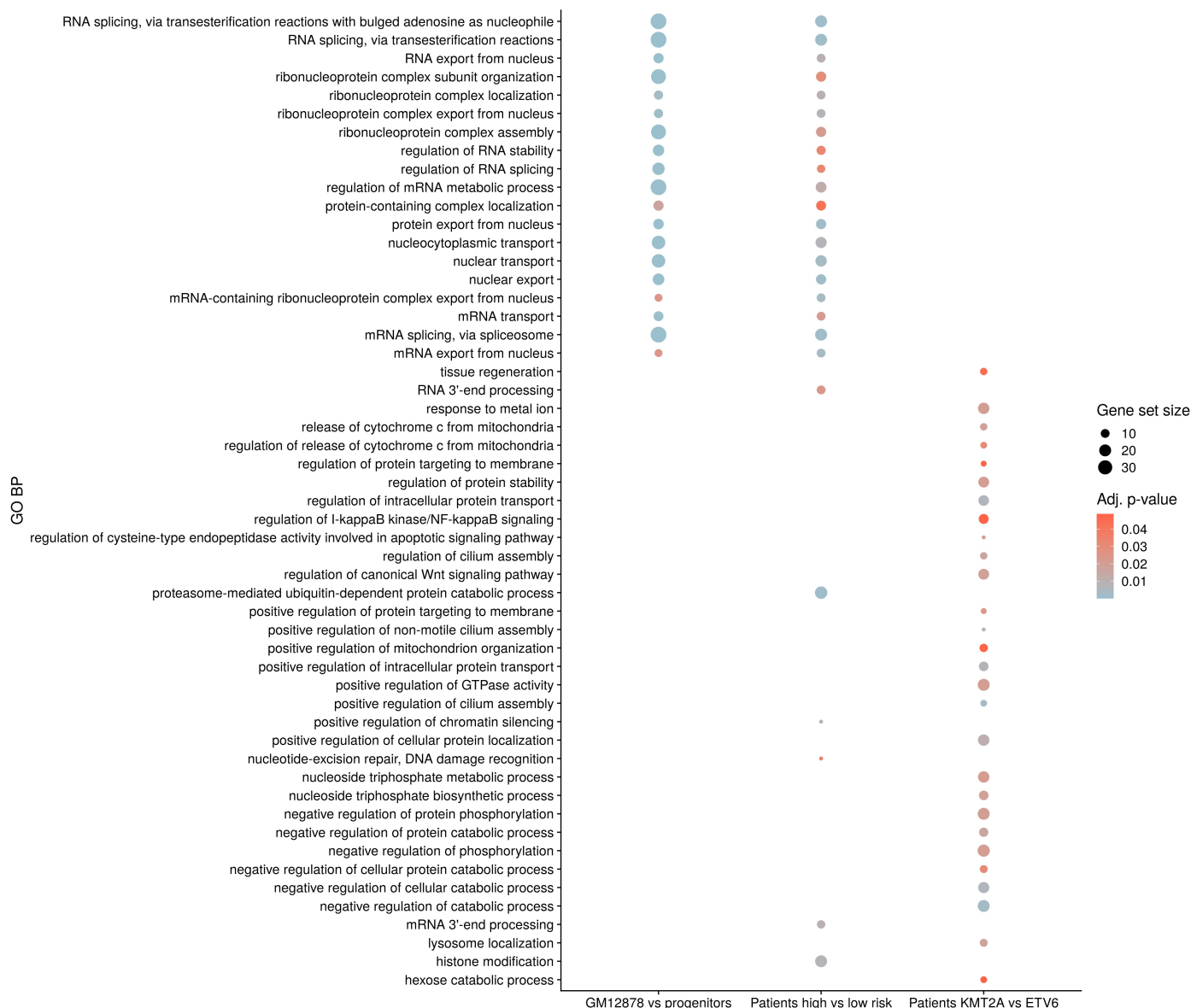
Supplementary Figure 22. Events co-occurring in high-risk patients from different fusion backgrounds. UpSet plot with the total number of events that co-occur with high risk associated with every fusion group. The bar plot indicates the total of number of events in each subset and the dots and lines the overlap of the events with every fusion group. Co-occurrence was measured as having average PSI = 0.5 or higher in two or more groups.



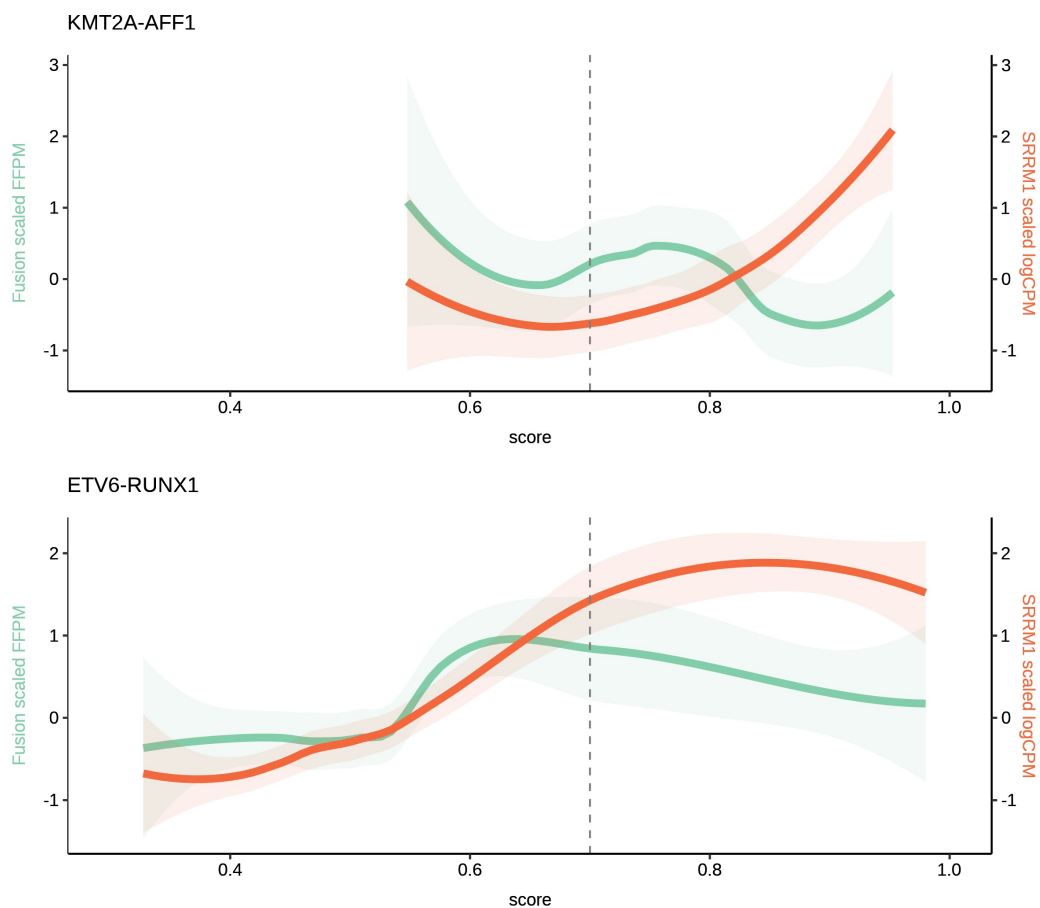
Supplementary Figure 23. EIF4H major isoforms expression distribution. Expression of the two most abundant EIF4H isoforms across patients separated by risk group. Isoform expression is given as $\log_{10}(\text{TPM} + 0.001)$ (y axis). The lines indicates the trend distribution for each isoform using smoothing with a linear model (method=*lm* in R).



Supplementary Figure 24. Predictor score boxplot by group of B-cell samples. Score grouped by B-ALL cell lines from CCLE, B-cell progenitors from CHOP cohort and set of 6 cell lines from GM12878, p-value obtained from a t-test mean comparison.



Supplementary Figure 25. Gene set enrichment analysis with splicing events. Pathway enrichment analysis using genes with splicing events that change significantly in any of 3 different comparisons: 1) GM12878 vs. B-cell progenitors, 2) high-risk vs. low-risk B-ALL patients, and 3) *KMT2A-r* vs *ETV6-r* patients.



Supplementary Figure 26. Score distribution in relation to fusion and SRRM1 expression. For the KMT2A-AFF1 (upper panel) and the ETV6-RUNX1 (lower panel) we represent on the x axis our signature score, on the left Y axis the fusion expression scaled and centred at 0 FFPM, and on the right Y axis the scaled and centred at 0 log2CPM expression of SRRM1.

Variable	Stats / Values	Freqs (% of Valid)	Missing(%)
Project	CHOP	18 (3.5%)	0 (0.0%)
	LUND	193 (37.8%)	
	PMJCI	50 (9.8%)	
	SJH	58 (11.4%)	
	TARGET_phase1	12 (2.4%)	
	TARGET_phase2	179 (35.1%)	
Type	ALL	510 (100%)	0 (0.0%)
Tissue	Blood	102 (20%)	0 (0.0%)
	Bone Marrow (BM)	408 (80%)	
Blasts	Mean (sd): 93.4 (6.8); IQR (CV): 8 (0.1); min < med < max: 43 < 95 < 100	—	404 (79.2%)
Time sample extraction	DX (diagnosis)	428 (83.9%)	0 (0.0%)
	RL (relapse)	82 (16.1%)	
Gender	Female (F)	240 (47.1%)	0 (0.0%)
	Male (M)	270 (52.9%)	
Age (month)	Mean (sd): 67.9 (62.8); IQR (CV): 96 (0.9); min < med < max: 0 < 48 < 364.9	—	2 (0.4%)
Fusions	None	239 (46.9%)	0 (0.0%)
	KMT2A-AFF1	59 (11.6%)	
	ETV6-RUNX1	56 (11%)	
	TCF3-PBX1	31 (6.1%)	
	KMT2A-MLLT3	18 (3.5%)	
	KMT2A-MLLT1	15 (2.9%)	
	P2RY8-CRLF2	9 (1.8%)	
	IGH-DUX4	6 (1.2%)	
	BCR-ABL1	4 (0.8%)	
	[57 others]	73 (14.3%)	
Cell of Origin	B Cell ALL	320 (62.7%)	0 (0.0%)
	B-Precursor	190 (37.3%)	
First event	None	31 (16.2%)	319 (62.5%)
	Relapse	160 (83.8%)	
Event Free Survival Time (days)	Mean (sd): 1144.8 (1046.1); IQR (CV): 683.5 (0.9); min < med < max: 77 < 851 < 4383	—	319 (62.5%)
Vital Status	Alive	80 (41.9%)	319 (62.5%)
	Dead	111 (58.1%)	
Overall Survival Time (days)	Mean (sd): 1810.7 (1230.1); IQR (CV): 2122 (0.7); min < med < max: 187 < 1290 < 4383	—	319 (62.5%)

Supplementary Table 1. Summary with the clinical information for the samples selected from the multiple cohort study used on the analysis. For the rows Blast, Age (month), Event Free Survival Time (days) and Overall Survival Time (days), we calculated a Mean, a standard deviation (sd), the interquartile range (IQR), coefficient of variation (cv), minimum value (min), median (med) and maximum value (max).

Gene name	beta	HR (95% CI for HR)	wald.test	p.value	Gene name	beta	HR (95% CI for HR)	wald.test	p.value
BAG2	0.75	2.1 (1.3-3.4)	11	0.012	NRIP2	0.32	1.4 (0.71-2.7)	2.8	0.42
BOP1	1	2.9 (1.7-4.7)	18	0.00039	NSUN3	0.9	2.5 (1.6-3.7)	21	0.00012
C8orf88	0.51	1.7 (1.1-2.5)	8	0.045	NSUN4	0.45	1.6 (0.99-2.5)	5.5	0.14
DHX33	0.35	1.4 (0.95-2.1)	4.9	0.18	ODC1	1	2.7 (1.8-4.1)	26	1.10E-05
DHX37	0.6	1.8 (1.2-2.7)	11	0.013	PABPC1	0.54	1.7 (1-2.9)	6.2	0.1
DPH6	0.54	1.7 (1.1-2.6)	8.2	0.043	PABPC4	0.71	2 (1.3-3.1)	13	0.004
EIF2AK4	0.47	1.6 (0.76-3.4)	3.5	0.32	PCBP3	0.95	2.6 (1.3-5)	9.9	0.019
EIF3C	0.8	2.2 (1.4-3.4)	15	0.0016	PPP1R15B	-0.25	0.78 (0.39-1.5)	2.5	0.48
EIF5	0.47	1.6 (0.88-2.9)	4.3	0.23	PPRC1	0.63	1.9 (1.1-3.3)	6.9	0.076
EIF5A2	0.88	2.4 (1.5-3.9)	14	0.0031	RBM15	0.43	1.5 (0.98-2.4)	5.5	0.14
EXOSC6	0.8	2.2 (1.5-3.3)	17	0.00071	RBM24	0.72	2.1 (1.3-3.2)	12	0.0069
FAM120A	0.79	2.2 (1.3-3.8)	9.9	0.02	RIOX2	0.91	2.5 (1.2-5.2)	8	0.047
FAM207A	0.9	2.5 (1.6-3.9)	18	5.00E-04	RNU2_1	0.86	2.4 (1-5.4)	6.3	0.099
GLUL	0.75	2.1 (1.4-3.3)	14	0.0034	RNU6_1	0.7	2 (1.2-3.3)	9.4	0.024
GNL3L	0.99	2.7 (1.3-5.6)	9.2	0.027	RRP15	0.82	2.3 (1.4-3.7)	13	0.0042
IGF2BP2	0.47	1.6 (1.1-2.4)	7.2	0.066	RRP8	0.69	2 (1.3-3)	12	0.007
KRI1	0.76	2.1 (1.2-3.8)	9.2	0.026	RRS1	1.4	4.2 (2.4-7.2)	28	3.80E-06
MBNL1	0.75	2.1 (1.1-4)	8	0.045	SAMD4A	0.74	2.1 (1.4-3.2)	14	0.0024
MTRF1L	1	2.8 (1.5-5.3)	12	0.0083	SLC29A2	0.71	2 (1.4-3)	14	0.0031
MYBBP1A	0.78	2.2 (1.3-3.7)	9.7	0.022	SORD	0.86	2.4 (1.4-3.9)	14	0.0035
MYC	1.2	3.2 (1.8-5.6)	19	0.00029	SRRM1	0.54	1.7 (1.2-2.6)	9.2	0.027
NAF1	-0.4	0.67 (0.35-1.3)	3.5	0.32	TCOF1	0.64	1.9 (1.1-3.2)	7.7	0.052
NAP1L1	0.62	1.9 (1.2-2.8)	11	0.01	TOP1MT	1.1	3.1 (1.9-5.1)	23	5.00E-05
NLE1	1.3	3.7 (2.3-5.9)	29	2.00E-06	UPF1	0.67	2 (0.9-4.2)	4.7	0.19
NOL6	0.91	2.5 (1.4-4.5)	11	0.012	URB1	-0.43	0.65 (0.34-1.3)	3.6	0.31
NOL9	0.79	2.2 (1.2-3.9)	9.2	0.026	URB2	0.69	2 (1.3-3.2)	11	0.014
NOM1	0.5	1.6 (1-2.6)	6.2	0.1	UTP20	-0.79	0.45 (0.24-0.84)	8	0.045
NOP14	0.78	2.2 (1.3-3.8)	9.6	0.023					

Supplementary Table 2. Summary table with the Cox proportional hazards regression results per gene adjusted by age and gender. Beta is the regression coefficient, a positive sign means that the hazard (risk of relapse) is higher, and thus the prognosis worse, for subjects with higher values of that variable. HR(95% CI for HR) is the hazard ratio with upper and lower 95% confidence interval.