Published online 9 December 2022

https://doi.org/10.1093/narcan/zcac041

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

Adria Closa[1,2,3], Marina Reixachs-Solé[1,2,3], Antonio C. Fuentes-Fayos[4,5,6], Katharina E. Hayer[7], Juan L. Melero[1,2,3], Fabienne R.S. Adriaanse[8], Romy S. Bos[8], Manuel Torres-Diz[7], Stephen P. Hunger[9], Kathryn G. Roberts[10], Charles G. Mullighan[10], Ronald W. Stam[8], Andrei Thomas-Tikhonenko[7,11], Justo P. Castaño[4,5,6,12], Raúl M. Luque[4,5,6,12] and Eduardo Eyras[1,2,3,13,14,*]

[1]The Shine-Dalgarno Centre for RNA Innovation, John Curtin School of Medical Research, Australian National University, Canberra, Australia, [2]Centre for Computational Biomedical Sciences, John Curtin School of Medical Research, Australian National University, Canberra, Australia, [3]EMBL Australia Partner Laboratory Network at the Australian National University, Canberra, Australia, [4]Maimonides Biomedical Research Institute of Cordoba (IMIBIC), Cordoba, Spain, [5]University of Cordoba (UCO), Cordoba, Spain, [6]Reina Sofía University Hospital, Cordoba, Spain, [7]Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, USA, [8]Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands, [9]Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, USA, [10]Department of Pathology, St. Jude Children's Research Hospital, Memphis, USA, [11]Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, USA, [12]Centro de Investigación Biomédica en Red de Fisiopatología de la Obesidad y Nutrición, (CIBERobn), Cordoba, Spain, [13]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain and [14]Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain

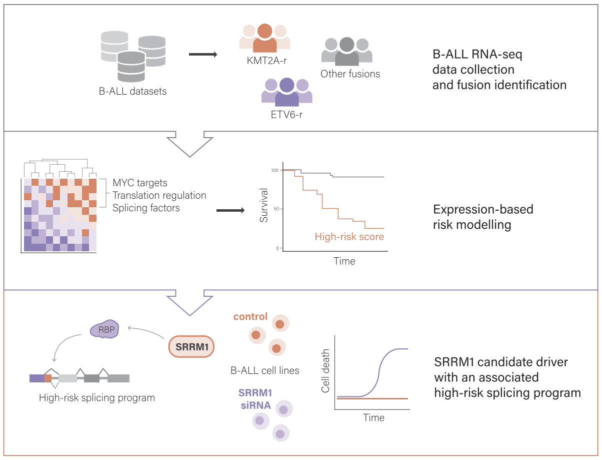
Received June 14, 2022; Revised October 09, 2022; Editorial Decision November 13, 2022; Accepted November 25, 2022

ABSTRACT

A significant proportion of infant B-cell acute lymphoblastic leukemia (B-ALL) patients remains with a dismal prognosis due to yet undetermined mechanisms. We performed a comprehensive multicohort analysis of gene expression, gene fusions, and RNA splicing alterations to uncover molecular signatures potentially linked to the observed poor outcome. We identified 87 fusions with significant allele frequency across patients and shared functional impacts, suggesting common mechanisms across fusions. We further identified a gene expression signature that predicts high risk independently of the gene fusion background and includes the upregulation of the splicing factor *SRRM1.* Experiments in B-ALL cell lines provided further evidence for the role of SRRM1 on cell survival, proliferation, and invasion. Supplementary analysis revealed that SRRM1 potentially modulates splicing events associated with poor outcomes through protein-protein interactions with other splicing factors. Our findings reveal a potential convergent mechanism of aberrant RNA processing that sustains a malignant phenotype independently of the underlying gene fusion and that could potentially complement current clinical strategies in infant B-ALL.

[*]To whom correspondence should be addressed. Tel: +61 2 6125 3225; Email: eduardo.eyras@anu.edu.au

GRAPHICAL ABSTRACT



INTRODUCTION

B-cell acute lymphoblastic leukemia (B-ALL) is the most common form of childhood cancer worldwide and one of the leading causes of cancer-related deaths in children (1). B-ALL presents a general lack of mutations in gene drivers that could be therapeutically targeted and are common in solid tumors and other leukemias (2). In contrast, B-ALL presents frequent chromosomal translocations that lead to the expression of gene fusions that are associated with marked differences in response to chemotherapy and survival. For instance, the frequent *ETV6-RUNX1* and *TCF3-PBX1* fusions have been associated with better prognosis (3), whereas *BCR-ABL1*, the highly frequent rearrangements of the gene *KMT2A* (*KMT2A-r*), and the less frequent *TCF3-HLF* are associated with poor prognosis (3,4). Other rarer, less studied fusions remain with an uncertain prognosis.

The prevalence of gene fusions has spurred multiple efforts to identify treatments that target them or their downstream effectors, albeit with limited success (5). In particular, many fusions involve transcription factors, which are difficult to target directly (6). Despite these challenges, combination chemotherapy and recent advances in Chimeric Antigen Receptor (CAR) T-cell therapy have led to a 90% increase in the 5-year survival rate in children younger than 15 years and a 75% increase for adolescents (15–19 years). However, infant B-ALL remains with a bad prognosis, especially for the *KMT2A-r* cases, which occur in about 80% of the infant patients during embryonic/fetal hematopoiesis (7). A total of 135 different translocation partners have been identified for *KMT2A*, with the most frequent ones being members of transcriptional elongation complexes, accounting for 90% of all *KMT2A-r* cases (8). Furthermore, many of these fusions may occur at any age and are also frequent in acute myeloid leukemia (AML) (8). Genome-sequencing studies of *KMT2A-r* B-ALL have confirmed a very low frequency of somatic mutations, suggesting that *KMT2A-r* may not require additional alterations to induce transformation (9,10). However, B-ALL cannot be recapitulated in pre-clinical models that only integrate the fusion, suggesting that additional alterations are necessary for leukemogenesis (11,12).

The rarity of many of the fusions in B-ALL and their apparent links to functionally distinct pathways complicate their interpretation and the identification of effective therapies. On the other hand, there is increasing evidence for convergent molecular signatures in B-ALL that indicate similar disease progression patterns and common therapeutic vulnerabilities, despite presenting different genetic alterations. For instance, a BCR-ABL1-like B-ALL subtype was described that shows a gene expression profile and a therapeutic vulnerability similar to *BCR-ABL1* patients, despite not presenting the *BCR-ABL1* fusion (13). For *KMT2A-r* fusions, the functional impacts of the different fusion gene pairs have been linked to common downstream mechanisms of chromatin and transcription dysregulation (12). The occurrence of *KMT2A-r* across different ages and lineages and their general association with poor prognosis suggests that a convergent phenotype might be potentially identified in B-ALL associated with poor prognosis. We hypothesized that this phenotype might be captured in the transcriptome and present in high-risk B-ALL tumors independently of the fusion background.

Here, we describe a comprehensive multi-cohort, infant, and child-focused characterization of high-risk B-ALL transcriptomics signatures. We identified a predictive gene expression signature of high-risk independent of the fusion background. This signature was mainly composed of ribosome biogenesis and RNA processing regulators, including the splicing factor *SRRM1*. Experiments in B-ALL cell lines provided evidence for the functional role of SRRM1 in cell survival, proliferation, and invasion. Furthermore, we found an alternative splicing program associated with the high-risk signature potentially mediated by splicing factors interacting with SRRM1. Our results provide a new layer of molecular variation that has remained undetected so far and represents a potential source of novel prognostic markers and therapeutic strategies in B-ALL.

MATERIALS AND METHODS

Data availability

All samples were obtained from various sources through controlled or public access. The series from St Jude Children's Research Hospital (SJH) (EGAS00001000246) (14) and Lund University (LUND) (EGAS00001001795) (15) were downloaded from the European Genome-phenome Archive (EGA) and the corresponding clinical information was obtained from the associated publications. The series from Children's Hospital of Philadelphia (CHOP) (GSE115656) was downloaded from Gene Expression Omnibus (GEO) (16). Samples from TARGET (Therapeutically Applicable Research to Generate Effective Treatments) were downloaded from the TARGET data portal at the National Cancer institute (NIH) together with the associated clinical information, corresponding to dbGAP accessions phs000463 (ALL phase 1) and phs000464 (ALL phase 2). Data from patients from the Princess Maxima Center for Pediatric Oncology (PMJCI) from (17) was obtained from the authors. We also analyzed data from 12 B-ALL cell lines coming from the Cancer Cell Line Encyclopedia (CCLE) (18), from normal blood and spleen samples from The Genotype-Tissue Expression (GTEx) project (19) and from 16 B-cell progenitors from CHOP (GSE115656) (16,20). We also analyzed an independent cohort of B-ALL patient RNA-seq datasets (21), with EGA IDs EGAD00001004461, EGAD00001006609 and EGAD00001007530, downloaded under Data Access Agreement between Children's Hospital of Philadelphia and St. Jude Children's Research Hospital – Washington University Pediatric Cancer Genome Project.

Clinical information and data extraction

All the information related to the clinical annotations and sample extraction is described in detail in their respective publication. A summary of the sequencing platforms used for each cohort included in this study is provided in Supplementary Table S1. A detailed description of the samples selected from each project with their relevant clinical information is provided in Supplementary Tables S2 and data file 1. We only used samples classified as B-cell ALL that had at least 25% of the reads mapping to the genome (GRCh38) and such that these corresponded to at least 5M reads. FASTQ files from SRA for the TARGET samples were extracted using the SRAToolKit (v 2.9.0) (https://github.com/ncbi/sra-tools). FastQC (22) was used for quality control of all the FASTQ files. All the samples were processed with the same pipeline outlined in Supplementary Figure S1.

Fusion detection

We used STAR-Fusion v.1.4.0 (23) to identify gene fusions from the RNA-seq data. The index was generated using the Gencode (v27) annotation and the GRCh38 assembly. STAR-Fusion was run for each FASTQ file using the default parameters described at https://github.com/STAR-Fusion/STAR-Fusion/wiki/Home/. We required at least one read count supporting the fusion junction given by the field JunctRC (or JunctionReads in the latest version of the manual) in the STAR-Fusion output, one read count connecting the fusion junction (SpanRC or SpanningFrags field), 0.1 Fusion Fragments Per Million total reads (FFPM), and junction reads that cover at least 25 bases on both sides of the breakpoint (indicated as 'YES_LDAS' in the STAR-Fusion output). The fusion allele frequency (FAF) was defined as the average of the allele frequency for both partners of the fusion pair, i.e. $\mathrm{FAF} = (\mathrm{FAF}_L + \mathrm{FAF}_R)/2$, where each value FAF_i, for $i = \mathrm{L}, \mathrm{R}$ was defined as $\mathrm{FAF}_i = \mathrm{F}_i/(\mathrm{F}_i + \mathrm{WT}_i)$, where the F_i represents the number of reads that support the fusion breakpoint and WT_i represents the number reads that support the wild type fragment of the gene, not present in the fusion.

Fusion filtering and classification

Fusion calls involving pseudogenes were removed from the output, as well as fusions between paralogous genes (genes with 70% or more sequence identity), as they were considered potential artifacts. Fusions between immunoglobulin or hemoglobulin genes were also discarded. Additional filters for possible false positives were applied: A promiscuity filter was used to remove fusions involving genes paired with more than one other gene within the same sample, known to be potential artifacts from the library preparation (24). Moreover, we removed all predicted fusions that occurred only in one project to avoid project biases. We also filtered out fusions previously detected in non-cancerous tissues or cells (25), detected in normal samples from TCGA (26), or seen with STAR-fusion in RNA-seq data from normal blood cell types (27).

Additionally, we only kept fusions that appeared in five or more patients. Fusions involving genes previously reported to have mutations or fusions in leukemia were kept independently of these filters, but only if they appeared in five or more patients. Finally, the FAF was used to select the most relevant fusions (see Methods section Calculation of Fusion Allele Frequency for details). Based on the FAF distribution across all patients from the different cohorts, each fusion was required to have a median FAF >0.1. The partner with the lowest allele frequency in the fusion was required to have a median of 0.01 for the individual FAF value.

We classified the fusions into four major groups: (i) 'ALL', which indicated those already reported in any of the analyzed ALL datasets or reported previously in the databases COSMIC (28), TCGA (26) or MitelmanDB (29); (ii) 'blood', which indicated those fusions known to appear in other hematological malignancies according to the same databases; (iii) 'solid tumors', which indicated fusions known in other solid tumors and present in the same public databases and (iv) 'novel', which indicated fusions that were not present in the public databases. Fusions were grouped and labeled according to the most recurrent partner for differential expression analysis, co-occurrence analysis, and visualization purposes. Fusion groups that were not showing a pattern of mutual exclusion with any of the other groups according to a waiting time model for mutually exclusive cancer alterations implemented by the R package TiMEx (30) were grouped as 'Other'.

Analysis of domains disrupted by fusions

The same fusion breakpoint given by the RNA-seq reads was used if this occurred inside an exon expressed in the fusion. Otherwise, the positions used were the last base of the last exon from gene 1 included in the fusion, and the first base of the first exon from gene 2 included in the fusion. Using these values, we defined the breakpoints for genes 1 and 2 of the identified fusions. PFAM domains mapping to the proteins encoded by each fusion gene were extracted from Biomart (31), the protein coordinates of the domain span were converted to genomic coordinates and overlapped with the fusion breakpoints of the corresponding genes to establish for each breakpoint whether the domain was kept or lost as a result of the fusion.

Gene expression and functional enrichment analysis

Transcript level quantification for the Gencode transcriptome release 27 (GRCH38.p10) (32) was obtained in transcripts per million (TPM) units using Salmon (v 0.7.2) (33). Gene level quantification was obtained by transforming transcript TPMs to counts per gene using the *tximport* library function from Bioconductor (34).

For differential expression analyses, we considered the patients with only one identified fusion (after performing all the filtering steps) and for the most frequent fusions: 87 patients for *KMT2A-r*, 68 for *ETV6-r*, 9 for BCR-r, 36 for P2RY8-r, 14 for PAX5-r, 25 for PWLC1-r, 5 for RUNX-r, 35 for ST3GAL1-r, 33 for TCF3r, 33 for TTYH3-r, and 12 for ZNF384-r, plus the 133 patients with no fusions detected. We calculated the differentially expressed genes between pairs of groups. The read counts per gene were transformed to $\log_2$ counts per million (logCPM) using edgeR (35), and genes with mean log CPM < 0 were filtered out. The data was normalized with the TMM method from the edgeR package. Differential expression analysis was performed with LIMMA (36) using the function *limma.voom* adjusted by SVA with the covariables of sex, project, and tissue (bone marrow or peripheral blood).

Gene set enrichment analysis was performed with GSEA (37) for the list of hallmarks and for the biological process ontology using the pre-ranked enrichment method, sorting all the genes by the value of $-\log_{10}(P\text{-value}) \cdot \log_2\text{FC}$ obtained from the differential expression analysis. In the case of splicing factors and RBPs, as there is no pathway or hallmark gene set associated with them in the available databases, we built a list of genes splicing factor and/or RBP function from previous studies (38,39) and run a preranked GSEA with the absolute value (Data file 2).

Gene selection to construct a predictive model of prognosis

We used gene expression data from 133 TARGET patients with complete clinical information about the age of diagnosis and time to the first relapse and calculated a $\log_2$ fold-change (log FC) per gene using the normalized log CPM mean expression between patients with relapse and without relapse. Similarly, we calculated a logFC from the gene expression of 140 patients from all other cohorts (SJH, LUND, CHOP and PMJCI) comparing the ones carrying only *KMT2A-r* against the ones with only *ETV6-r*. We only considered genes with a log FC > 0.5 in both comparisons and that were included in at least one of three sets: (i) the GSEA hallmark as MYC targets (v1 and v2), in (ii) the Gene Ontology Biological Process of translation (initiation, elongation, or termination) or (iii) and a list of genes encoding splicing factors and RBPs (Data file 2). For every gene, we applied a cox regression survival model adjusted by age and gender, selecting only the genes with a *P*-value <0.05 according to a Wald test (Supplementary Table S3). This produced a total of 39 overexpressed genes associated with prognosis and with our target biological functions and pathways. From these 39 genes, the expression variability between cohorts was evaluated using mean logCPM expression in each dataset and the maximum logFC between datasets. Genes with the highest variability across datasets (max log FC > 3) were removed to avoid any dataset-related bias, obtaining the final 37 genes to build the predictive model. Training data was restricted to the 133 TARGET B-ALL patients. The model consisted of a random forest, implemented using the randomForest library in R, with a total of 400 trees and four variables randomly sampled as a candidate at each split. The leave-one-out strategy was used to evaluate the prediction accuracy, while avoiding overfitting. A numeric k-score between 0–1 obtained from the prediction of the random forest model was used to classify the patients according to risk. A threshold was established on 0.7, according to accuracy measures, to classify the patients as high-risk (k-score ≥ 0.7) or low-risk (*k*-score < 0.7). To apply the predictive model to expression values given in terms of regularized-log (rlog) values, we used the TARGET samples as before, applying a regularized log transformation using the package DESeq2 with the *rlog* function (40). We then built and tested the model as before, using the rlog values instead of the log CPM values.

Differential splicing analysis

SUPPA (41,42) was used to perform the differential splicing analysis. SUPPA predicts the relative inclusion and differential splicing of the events using isoform-level relative abundances. Using RT-PCR experiments, SUPPA's accuracy was shown to be comparable to methods based on the direct quantification of event PSI from reads (41,42). SUPPA *generateEvents* was used to generate alternative splicing events defined from protein-coding transcripts and covering the annotated ORFs. The relative inclusion of each event was calculated as a Percent Spliced In (PSI) value with SUPPA *psiPerEvent* using the transcript abundances in TPM units obtained before. A minimum total expression of the transcripts involved in the event of 1 TPM was required. Events without a defined PSI value in more than 10% of the patients across all cohorts were discarded. These included events that did not pass the transcript expression filter or that had all the transcripts involved in the event with zero expression. The remaining missing PSI values were imputed using nearest neighbor averaging with the *impute.knn* function in R from the Impute library (43). To test the significant differential inclusion of the events in the comparisons of high against low-risk patients and in *KTM2A*-r against *ETV6*-r patients, a ΔPSI was calculated as the difference of the mean PSI from each group. We discarded all events with a standard deviation (SD) across groups lower than 0.1. We applied a linear regression model with a logit transformation of the PSI to estimate the significance of the splicing changes and adjusted the *P*-value by calculating a false discovery rate (FDR), using the same covariables adjustment as in the differential expression analysis described previously. We considered significant all the changes with $|\Delta\text{PSI}| > 0.2$ and an FDR corrected *P*-value < 0.01. Differential splicing between B-cell precursors and GM12878 was calculated with the SUPPA *diffSplice* command with default options.

Motif enrichment analysis

We searched for RBP binding motifs on the regions neighboring each splicing event with MoSEA (https://github.com/comprna/MoSEA) (39). MoSEA was run against a database of Position Frequency Matrices (PFM) and *k*-mers (6-mers) associated with each RBP. Enrichment was assessed by comparing a set of events differentially spliced between conditions with a set of events with no significant change between the same conditions. For each motif, MoSEA calculated a z-score from the comparison of the observed frequency observed in differentially spliced events with the distribution of frequencies in 100 control subsamples of the same size, considering the length distribution and GC content of the differentially spliced events set. We considered those motifs PFMs and 6-mers with *z*-score > 1.5.

Retrieving protein-protein interactions

We used the STRING database (44) to retrieve protein–protein interactions with the detailed scores of the links between proteins. Only those with experimental scores different from 0 and a combined score higher than 900 were kept.

Co-occurrence analysis of differentially spliced events and high risk

An alternative exon was considered included for PSI > 0.5 and absent otherwise for each sample. For every fusion group, a matrix was built with the presence or absence of events in each patient. The co-occurrence of the events with high risk was tested with a probabilistic model of species co-occurrence implemented in the R package co-occur (45).

Functional enrichment analysis of differentially spliced genes

Genes associated with differentially spliced events were tested for functional enrichment of Gene Ontology Biological Process terms with the R package clusterProfiler from Bioconductor (46). Benjamini–Hochberg (BH) correction was used to calculate adjusted *P*-values (*q*-values). Only ontologies with *P*-value and *q*-value <0.05 were selected.

Leukemia cell lines selection

SEM, MHH-CALL-3, KOPN-8, NALM-19, REH and SUP-B15 cells were obtained from the Leibniz Institute DSMZ (#ACC546, #ACC339, #ACC552, #ACC522, #ACC22 and #ACC389, respectively) and cultured according to the supplier's recommendations. These cell lines were previously checked for mycoplasma contamination by PCR as previously reported (47). Results were expressed as a percentage with respect to scramble-transfected controls.

RNA isolation, real-time qPCR, and customized qPCR dynamic array based on microfluidic technology

Total RNA from leukemia cell lines was extracted with TRIzol® Reagent (ThermoFisher Scientific, #15596026). Total RNA concentration and purity were assessed by Nanodrop One Microvolume UV-Vis Spectrophotometer (ThermoFisher Scientific). For qPCR analyses, total RNA was retrotranscribed by using random hexamer primers and the RevertAid RT Reverse Transcription Kit (ThermoFisher Scientific, #K1691). Thermal profile and qPCR analysis to obtain absolute mRNA copy number/50 ng of sample of selected genes are reported elsewhere (48). To control the possible variations in the efficiency of the retrotranscription reaction, mRNA copy numbers of the different transcripts analyzed were adjusted by *ACTB* expression. Specific primers for human and mouse transcripts including *ACTB* and *SRRM1* genes were specifically designed with the Primer3 software [*SRRM1* (NM_001303448.1)—forward: GTAGCCCAAGAAGACGCAAA, reverse: TGGTTCTGTGACGGGGAG; *ACTB* (NM_001101)—forward: ACTCTTCCAGCCTTCCTTCCT, reverse: CAGTGATCTCCTTCTGCATCCT].

Silencing of splicing factors by specific small interfering RNA

Pre-designed and validated specific small interfering RNA (siRNA) oligos for knockdown of endogenous *SRRM1* (#s20018; Silencer® Select siRNAs; ThermoFisher Scientific) were used, which is a pre-validated siRNA. Briefly, cells (n = 500 000 cells/well) were transfected with 25 nM of each siRNA individually using Lipofectamine® 3000 Transfection Reagent (ThermoFisher Scientific, # L3000075) according to the manufacturer's instructions. Silencer® Select Negative Control siRNA (ThermoFisher Scientific, #4390843) was used as a scramble control. After 24 h, cells were collected for validation of the transfection by qPCR and seeded for different functional assays.

Proliferation rate determination

Cell proliferation in response to *SRRM1* silencing in leukemia cell lines was analyzed using the alamarBlue™ assay (Biosource International, Camarillo, CA, USA), as previously reported (49). Briefly, cells were seeded in 96-well plates at a density of 25 000 cells/well and serum-starved for 24 h. Then, proliferation was evaluated every 24 h using the FlexStation-III system (Molecular Devices, Sunnyvale, CA, USA) for up to 72 h. Results were expressed as a percentage referred to as scramble-transfected controls.

Apoptosis measurement

Apoptosis induction in response to *SRRM1* silencing in leukemia cell lines (25 000 cells/well onto white-walled multiwell luminometer plates) was performed by using Caspase-Glo® 3/7 Assay (Promega Corporation, #G8091) as previously reported (49). Briefly, Cells were seeded in 96-well white polystyrene microplate flat bottom clearplates at a density of 25 000 cells/well and serum-starved for 24h. Results were expressed as a percentage referred to as scramble-transfected controls.

Invasion rate determination

The invasion rate in response to *SRRM1* silencing in leukemia cell lines was assessed by using the 96-well cell Trans-well invasion assay (Basement Membrane-8 μm, AssayGenie, #BN01086) according to the manufacturer's protocol. The top chamber membrane was coated with the basement membrane solution. Afterward, cells were seeded in the top chamber in 0.5% serum media. Then, 10% FBS media was placed in the bottom chamber to promote cell invasion in all experimental conditions. Then, proliferation was evaluated every 24 h using the FlexStation-III system (Molecular Devices, Sunnyvale, CA, USA).

Statistical analysis for the B-ALL cell line experiments

Numerical results were evaluated for statistical differences by t-test, multiple *t*-tests, and two-way ANOVA. All statistical analyses were performed using Prism software v.9.0 (GraphPad Software, La Jolla, CA, USA). *P*-values < 0.05 were considered statistically significant. The plotted data represent the median (interquartile range) or means ± standard error of the mean (SEM). Significant difference from control conditions was indicated as * *P*-value < 0.05, ** *P*-value < 0.01, *** *P*-value < 0.001.

RESULTS

RNA sequencing identifies novel gene fusions in B-ALL

We collected RNA sequencing (RNA-seq) from 428 patient samples obtained at diagnosis from five different B-ALL studies (14–17,50) (Supplementary Table S2 and Data file 1). Patients' age distribution peaked around one year old, with most cases being classified as infants or young children (Figure 1A). We applied a comprehensive pipeline to study gene fusions, expression, and RNA splicing (Supplementary Figure S1) (Methods). To identify fusions likely to be associated with the B-ALL phenotype, we removed fusion candidates that we had also detected in non-cancer tissues and normal hematopoietic cells, as well as potential artifacts. We kept all fusion candidates reported in the clinical data of the studied cohorts or involving genes with acute leukemia mutations in COSMIC (28). Additionally, only candidates appearing in five or more samples and across different projects were considered. Starting from 1825 unique candidate fusions, these filters resulted in 158 unique high-confidence fusions (Supplementary Figure S2) (see Materials and Methods for details).

To prioritize the relevance of the identified fusions, we further calculated a fusion allele frequency (FAF) (23), defined as the proportion of the gene expression corresponding to a fusion averaged over the two genes participating in the fusion, and which represents a proxy for the fusion clonality (Supplementary Figure S3). We then considered only those fusions with a median FAF of at least 0.1 and having none of its genes with a median individual FAF <0.1. These analyses resulted in our final list of 87 different fusions (Data file 3) (see Materials and Methods for details).

Using these fusions, all five analyzed cohorts presented a similar distribution of fusions per patient (Supplementary Figure S4). Our analyses of the RNA-seq data recovered 81.35% of the most frequent fusions detected in the same samples with independent experimental methods (Supplementary Figure S5a and b). Although most of the identified fusions had been observed previously in ALL, we also identified fusions that had been reported before in other blood cancers or in solid tumors, as well as novel fusions (Supplementary Figure S5c). Moreover, we identified known B-ALL fusions in 43 patients that did not have any fusion annotated in the published clinical information and determined the fusion partner in 7 cases that were only annotated as *KMT2A-r* (Data file 3).

We found that some of the identified fusions are overrepresented in specific age groups (Figure 1B). We recovered the known enrichment of *KMT2A-r* in infant cases and *ETV6-r* in children and young adults (1–18 years). Fusion groups such as *BCR-r* and *P2RY8-r* presented a bimodal or extended age distribution, including infant and child. Similarly, *PAX5-r* appeared in infants and at the upper extreme of childhood cases (Supplementary Figure S6). Despite these associations, age distributions and fusion frequencies may not represent the actual distribution in the population of leukemia patients, as there may have been sample collection biases in each cohort.

Overall, high-confidence fusions were detected for 70% of the samples at diagnosis. Grouping the fusions by the most frequent gene in the fusion pairs, we could identify six major fusion groups: *KMT2A-r* (20%), *ETV6-r* (16%), *ST3GAL1-r* (8%), *P2RY8-r* (8%), *TCF3-r* (8%) and *TTYH3-r* (7%) (Figure 1C). In the cohorts analyzed, we also found *PAX5-r* (3%), *ZNF384-r* (3%), *BRC-r* (2%), *RUNX1-r* (2%) and *GSE1-r* (2%), as well as a group of low-frequency fusions that appeared in 18% of the patients. The calculated fusions in diagnostic samples showed a clear pattern of mutual exclusions and a frequent co-occurrence of the inverse fusion for *KMT2A-r*, *ETV6-r*, *PAX5-r* and *BCR-r*, but not for other fusions (Figure 1c). Apart from known fusions in B-ALL and hematological malignancies we described fusions previously observed in solid tumors, *GSE1-SLC7A5* and *CBFA2T3-PIEZO1*. Among the fusions that have not been previously described in cancer, the majority can be attributed to new partners of already known fusion genes in B-ALL or other hematological malignancies such as *ST3GAL1* or *P2RY8* (Figure 1C) (Data file 3).

Different gene fusions impact similar functional pathways

We calculated the breakpoints for each of the fusions detected from RNA-seq reads, either as the exon position where the fusion breakpoint was found or using the boundaries of the exons flanking the intron where the breakpoint was assumed to fall. Fusion genes presented multiple breakpoints (Figure 2A) (Supplementary Figure S7). Moreover, these breakpoints appeared in positions that potentially disrupted the protein domain content. This raised the question of whether different breakpoints in the same or different fusion genes may lead to similar functional impacts. To determine this, we calculated the domains that are maintained or lost in each fusion gene according to the identified breakpoints. In terms of the specific domains kept or lost in a fusion, we observed little overlap between fusions (Supplementary Figure S8). However, when we grouped the protein domains according to their functional ontologies, multiple similarities appeared, such as the loss of DNA-binding domains in *KMT2A-r* and *ETV6-r* and the loss of signal transduction domains in *P2RY8-r* and *BCR-r* (Figure 2B).

Common patterns of gene expression across diverse gene fusion backgrounds

To further characterize the cellular processes associated with the identified fusion groups, we calculated the differential expression patterns among patient groups. Despite the heterogeneity of samples used in the comparison, there were many significant expression changes associated with *KMT2A-r, ETV6-r*, *ST3GAL1-r*, *ZNF384-r* and *TCF3-r*

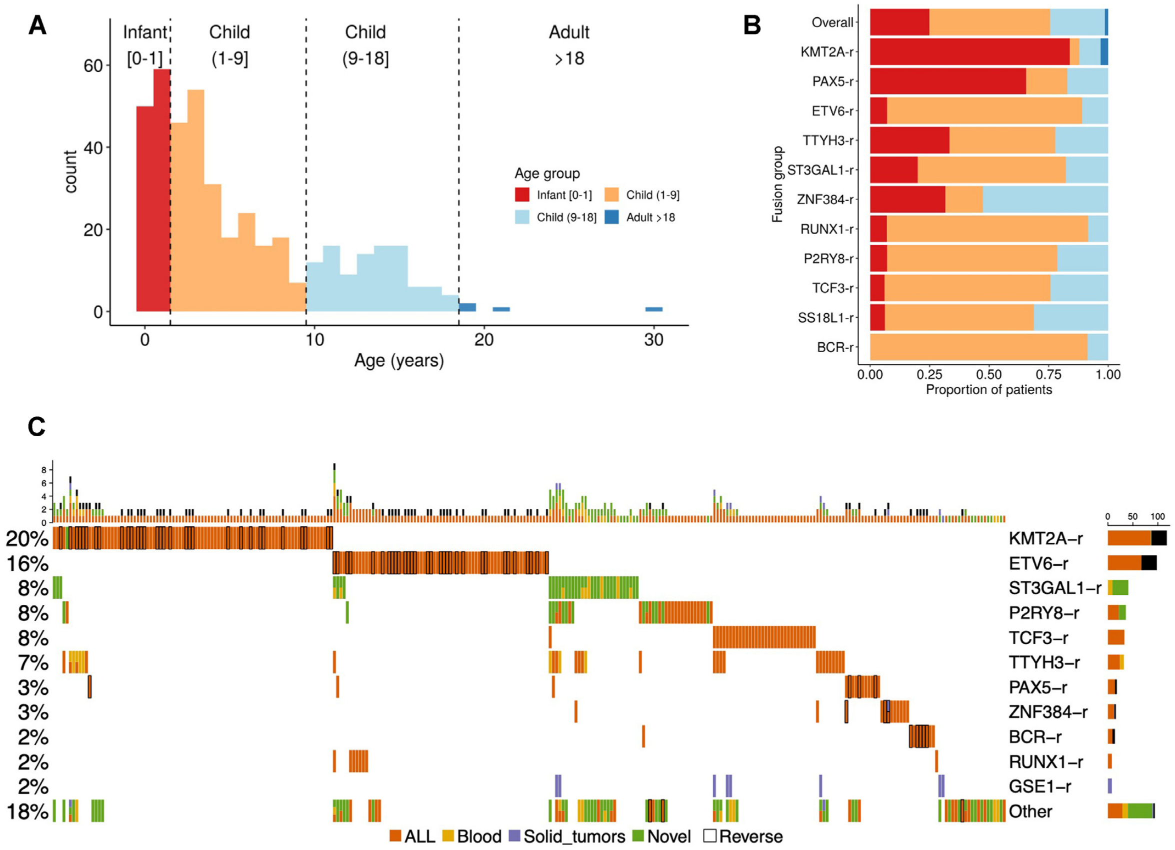


Figure 1. Multicohort identification of gene fusions in B-ALL. (**A**) Age distribution of the B-ALL patients studied. (**B**) The proportion of the gene fusion groups in each age group from (a). (**C**) Fusion oncoprint. The plot shows the most frequent fusions (rows) detected in patients (columns). Fusions are grouped by the most frequent gene in the fusion gene pairs. The bar plot above shows the number of fusions detected per patient. The specific gene fusion pairs in each patient are given in (Data file 3). A black border line indicates that the reverse fusion was identified in that patient. Cell colors indicate whether the fusions were detected before in B-ALL (orange), in a blood cancer (yellow), in a solid tumor (purple), or whether it was not reported in any cancer before (green). A patient with two different fusions detected in the same gene is depicted with two colors, and they were only kept if they had been previously reported in a tumor and occurred in at least five patients.

in comparison with the other fusion groups and with patients without fusions (Supplementary Figure S9) (Data file 4). These patterns included the HOXA overexpression characteristic of *KMT2A-r* patients (Supplementary Figure S10) (17). Interestingly, the differentially expressed genes associated with each fusion group did not generally overlap (Supplementary Figure S11), except for the *KMT2A-r* and *ETV6-r* groups, which showed opposite expression patterns (Supplementary Figure S12). This suggested that the functional alterations specific to *KMT2A-r* are reversed in *ETV6-r* tumors.

To investigate this possibility, we studied the pathways enriched or depleted in each fusion group. Each group tended to cluster independently, except for *ST3GAL1-r*, *PQLC1-r* and *TTYH3-r*, which presented similar pathway enrichments, and *P2RY8-r*, *RUNX1-r* and *PAX5-r*, which were similar to the cases with no fusions (Supplementary Figure S13). Moreover, genes overexpressed in KMT2A-r patients were strongly enriched in MYC targets, ribosome biogenesis, and RNA processing, including splicing and translation regulation (Supplementary Figure S13). Interestingly, MYC targets were upregulated in *KMT2A-r* patients but depleted in *ETV6-r* (Supplementary Figure S13), and transforming growth factor beta (TGF-b) signaling, which antagonizes MYC (51), was depleted in *KMT2A-r* patients. Furthermore, *MYC* expression was higher in *KMT2A-r* patients relative to the other patient groups and normal fetal-liver B-cells (Supplementary Figure S14), whereas *ETV6-r* patients showed *MYC* expression below normal fetal liver B-cells (Supplementary Figure S14). This suggested a gene expression pattern linked to *MYC*, in association with *KMT2A-r*, and reversed in *ETV6-r*.

A gene expression signature of high-risk B-ALL independent of the gene fusion background

B-ALL cases with *KMT2A-r* are generally considered to have a poor prognosis, whereas *ETV6-r* are deemed to be of good prognosis (3). We thus reasoned that the observed opposite expression patterns might be linked to these risk phenotypes. To investigate this possibility, we calculated the genes differentially expressed between patients harboring only *KMT2A-r* or *ETV6-r,* without any other detected fusion (87 *KMT2A-r* and 68 *ETV6-r*). All RNA-seq used was obtained at the diagnostic stage before treatment. Gene set enrichment analysis on the 1405 differen-

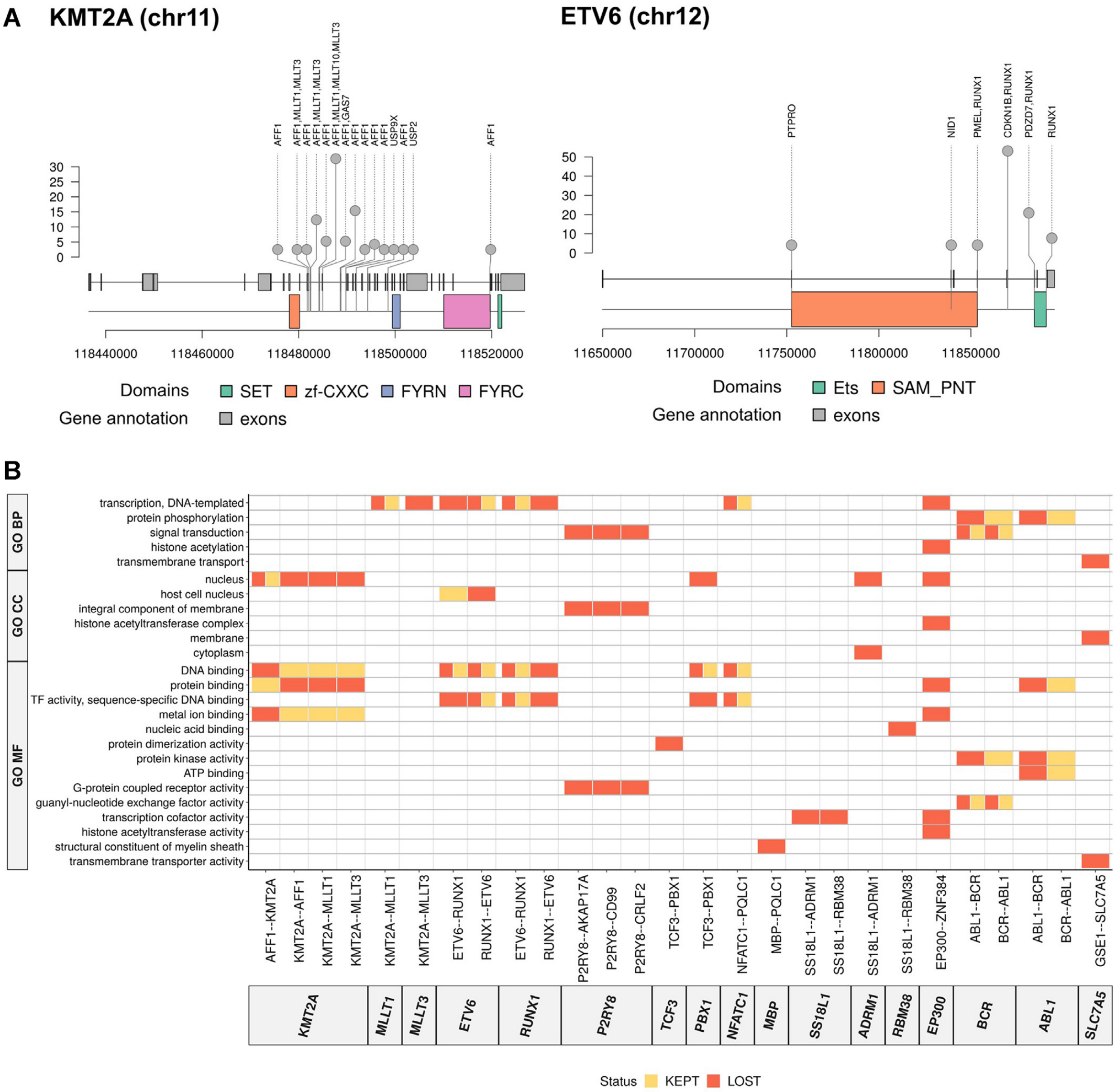


Figure 2. Functional impact assessment of gene fusions. (**A**) Identified breakpoints in the fusions involving *KMT2A* (left panel) and *ETV6* (right panel). Along the genomic locus (x-axis), we indicate the detected breakpoint's position, defined by an exonic position if the breakpoint occurred inside an exon, or by the last base of the last exon included in the fusion otherwise. The plot also shows the span of the functional domains in genomic coordinates. (**B**) Domains kept (orange) and lost (red) through gene fusions (x-axis) labeled according to gene ontologies (GO) (y-axis): biological process (BP), cellular compartment (CC) and molecular function (MF). The specific functional domains kept or lost in these fusions are given in Supplementary Figure S8.

tially expressed genes identified (FDR < 0.01 and |log FC| > 1) (Figure 3A) (Data file 5) revealed two main pathways, one associated with immune response and another related to MYC, translation regulation, and ribosome biogenesis (Figure 3B). As RNA processing is tightly controlled during normal hematopoiesis and is commonly dysregulated in hematological malignancies (52–54), we hypothesized that genes related to these pathways may separate patients according to their risk of relapse. Feature selection was conducted to extract a list of 39 genes from the target pathways showing overexpression in *KMT2A*-r. *RBM24* and *RNU6-1* were discarded from these genes as they showed high variability across the studied cohorts, resulting in a final signature of 37 genes (Figure 3C) (Supplementary Figure S15).

We trained and tested a predictive model for risk using the 133 patients from TARGET, a high-risk cohort with clinical follow-up information (Data file 1), using as endpoint the event-free survival and as the prognostic event the first relapse. Using a random forest leave-one-out cross-validation combined with a cox-regression, we obtained a significant separation of patients according to the risk of relapse (log-rank test *P*-value < 0.0001) (Figure 3D) and overall accuracy of 0.72, measured as the area under the receiver operating characteristic (ROC) curve (AUC) (Figure 3E). Importantly, even though the genes in our model were selected from the comparison of *KMT2A-r* and *ETV6-r* patients, only 11% of the 133 patients had KMT2A-r or ETV-6 fusions. Based on the observed values of sensitivity, specificity, and accuracy (Supplementary Figure S16a), we chose

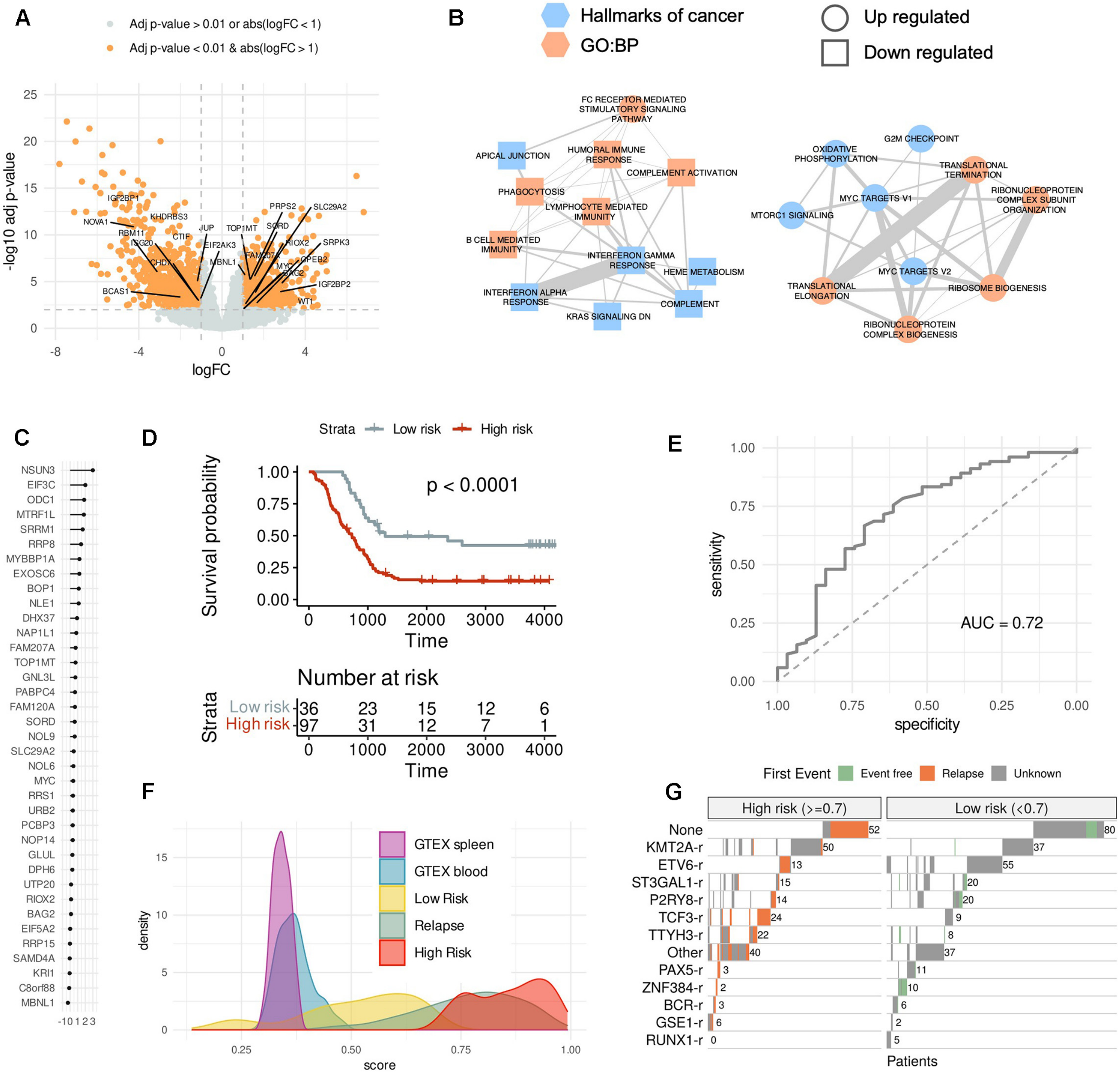


Figure 3. Gene expression signature of high risk. (**A**) Differential gene expression between a subset of patients with only *KMT2A-r* or only *ETV6-r*. The volcano plot shows for each gene the log2(fold-change) (x-axis) and the -log10(corrected p-value) (y-axis). We indicate the genes involved in RNA processing, RNA translation, and Ribosome biogenesis. (**B**) Enriched Cancer Hallmarks and Biological Process Gene Ontologies (GO:BP) of the genes significantly up or down-regulated in the comparison in (a). (**C**) Ranking of the 37 genes of the predictive model of high risk according to their relevance in the leave-one-out test (x-axis). Relevance is defined as the median of the accuracy per gene. (**D**) Kaplan–Meyer 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. (**E**) Average receiving operating characteristic (ROC) curve and area under the ROC curve (AUC) from the leave-one-out test for classifying patients into low and high risk. The ROC curve shows the specificity (x-axis) and sensitivity (y-axis) for the entire range of possible model score threshold values. Sensitivity is defined as the proportion of high-risk cases that are correctly predicted. In contrast, specificity is defined as the proportion of low-risk cases correctly predicted as low risk. (**F**) Risk score distribution for various sample groups: diagnostic samples predicted as high risk, diagnostic samples predicted as low risk, samples obtained at relapse, and blood and spleen samples from GTEX. (**G**) Classification of all the B-ALL diagnostic samples from each fusion subgroup (y-axis) into high (left) or low (right) risk according to our risk score. Patients with a clinical record indicating that they had relapsed are indicated in orange, whereas patients annotated with no relapse are indicated in green (event free). Patients with no clinical information are indicated in grey (unknown).

the decision boundary at a score 0.7 (≥0.7 for high risk and <0.7 for low risk). However, the same accuracy values were maintained with score thresholds between 0.5 and 0.7 (Supplementary Figure S16a).

To further evaluate our predictive signature, we trained a single model with all the 133 TARGET samples (Supplementary Figure S16b). We applied this model to normal blood and spleen samples from GTEX, all other B-ALL samples, and to 82 additional samples obtained at relapse, none of which were included in any of the analyses above. Remarkably, samples obtained at relapse showed a distribution similar to the high-risk samples and higher than the normal samples (Figure 3F). Furthermore, this model separated TARGET patients with and without relapse and detected other high-risk patients from the other cohorts (Figure 3G) (Supplementary Figure S16c). Our predictions showed that patients with fusions *KMT2A-AFF1*, *AFF1-KMT2A* and *TCF3-PBX1* were more frequently in the high-risk group. In contrast, patients with *ETV6-RUNX1*, *RUNX1-ETV6* and *ABL1-BCR* were more often classified as low risk, including the ability to stratify patients with higher risk inside every group of fusion (Figure 3G) (Supplementary Figure S16c) (Data file 6).

We further applied our model to an independent cohort of 188 B-ALL patients (21). Since the gene expression values for this dataset were only available in terms of regularized log (rlog) values, we rebuilt our model using the rlog expression values calculated from the TARGET samples. Using a leave-one-out cross-validation, this model maintained the same classification power as the previous model built with logCPM values (Supplementary Figure S17a), and the scores given by both models showed a high correlation ($R = 1$, P-value < 2.2e–16) (Supplementary Figure S17b). To test our predictions, we used the clinical classification of patients published by the independent study into three different risk groups (high, standard, and low) (21). Our model separated these three groups into significantly different score distributions, with the high-risk group showing the highest scores and the low-risk group showing the lowest signature scores (Supplementary Figure S17c). These analyses provide strong support for the ability of our model to predict the potential for relapse in B-ALL independently of the gene-fusion background. To make this model readily available to evaluate the risk on new sets of patients from expression data and to explore the features of the TARGET cohort, we integrated the predictive model into an interactive web resource available at https://github.com/comprna/risk_model_app.

SRRM1 as a candidate driver of progression and poor prognosis in B-ALL

One of the genes with the highest predictive value in our risk-prediction signature was *SRRM1*, which showed the highest correlation with the risk score (Pearson $r = 0.51$, P-value = 8.18e–30) (Data file 6) and has been associated before with poor prognosis in prostate cancer (55). Confirming this predictive power, patient samples classified as high-risk and relapse samples showed higher *SRRM1* expression than low-risk samples and normal samples from spleen and blood (Figure 4A). To further understand the functional transformations associated with *SRRM1* expression, we analyzed multiple samples from progenitor and mature B-cells. *SRRM1* was significantly downregulated in mature B-cells compared with B-cells progenitors (Figure 4B). Consistent with this, the risk score was significantly higher (0.50–0.73) in undifferentiated B-cells compared to differentiated B-cells (0.38–0.46; P-value < 0.001; Supplementary Figure S18; Data file 7). These results suggest a relevant role for *SRRM1* driving a highly proliferative phenotype.

SRRM1 has been described as an essential gene, but a knockdown is known to produce a reduction of the cell viability without killing the cell (Supplementary Figure S19), opening the door to *SRRM1* expression modulation as a potential strategy to reduce the aggressive phenotype of leukemia cells. To test this strategy, we determined proliferation, apoptosis, and invasion rate, after silencing *SRRM1* in six human B-ALL leukemia cell line models bearing distinct functional and phenotypic features: SEM, MHH-CALL-3, KOPN-8, NALM-19, REH and SUP-B15. The silencing of *SRRM1* expression was successful in all cell models (Figure 4C) and resulted in a significant decrease in proliferation rate in a time-dependent and cell line-dependent manner (Figure 4D). Furthermore, using the capase3/7 assay revealed that *SRRM1* silencing significantly induced apoptosis in all human leukemia cells (Figure 4E). Moreover, using a trans-well assay to evaluate the invasion capacity revealed that *SRRM1* silencing could potentially impair the capacity of these cells to invade surrounding tissues in all cell lines tested (Figure 4F). *SRRM1* expression analysis revealed that the cell lines had different basal expression patterns (from high to low levels: KOPN-8>SEM>REH>MHH-CALL-3>NALM-19>SUP-B15) and presented differences in the basal proliferation (Supplementary Figure S20). Furthermore, the proliferation rate with the siRNA at 48h was inversely correlated with the *SRRM1* basal expression and the silencing effectiveness (Supplementary Figure S20), supporting a tight association of *SRRM1* expression with this essential cellular function.

A candidate SRRM1-dependent splicing program associated with high-risk B-ALL

SRRM1 is a serine-arginine rich factor that is associated with splicing complexes and affects splicing through interactions with SR proteins (56) (Figure 5A). We thus decided to study the alternative splicing events potentially linked with SRRM1 and their association with poor prognosis. Analysis of differential RNA processing events between patients predicted as high and low risk identified a total of 422 events, with 342 of them affecting coding transcripts and showing enrichment of alternative first (AF) and skipping exons (SE) (Figure 5B; Data file 8). Those 342 events occurred in 271 genes, none of which were differentially expressed between the same conditions. expressed genes. Moreover, out of the 342 events with differential splicing, only 27 (7.8%) showed a correlation higher than 0.5 (Pearson R) between their PSI and the expression level of the host gene across the samples. This indicates that the splicing modulation identified is largely independent of the expression changes. Moreover, these significant events separated

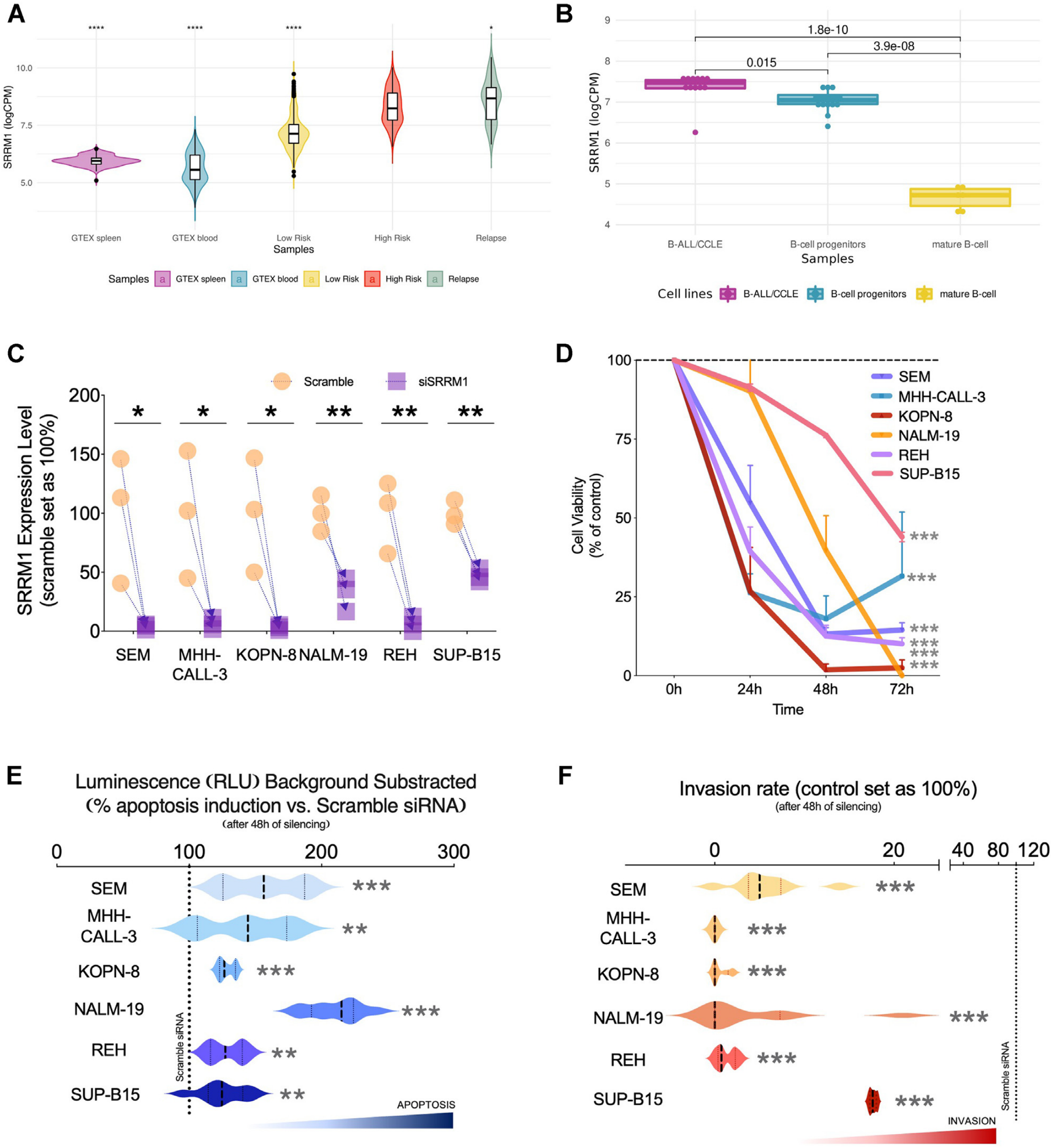


Figure 4. SRRM1 as a potential major driver of high-risk B-ALL. (**A**) Distribution of expression values for SRRM1 in log10 counts per million (logCPM) units (y-axis) in various sample groups: B-ALL diagnostic samples predicted to be high risk (score > 0.7), B-ALL diagnostic samples predicted to be low risk (score < 0.7), B-ALL relapse samples, and blood and spleen samples from GTEX (t-test applied to compare all the groups versus the high-risk group, **** for $P < 2.2e–16$, * for $P = 0.023$). (**B**) Normalized $\log_2$CPM expression of *SRRM1* in B-ALL cell lines from the cancer cell line encyclopedia (CCLE), B-cell progenitors, and a set of GM12878 biological replicates. *P*-values were obtained from a t-test mean comparison. (**C**) Validation of *SRRM1* silencing (y-axis) in cells (x-axis) by qPCR ($n = 3$). (**D**) Proliferation rate (y-axis) in response to *SRRM1* silencing in leukemia cell lines at different time points (x-axis) ($n = 3$ per time point). (**E**) Apoptosis rate (x-axis) in response to *SRRM1* silencing in leukemia cell lines (y-axis) ($n = 3$). (**F**) Invasion rate (x-axis) in response to *SRRM1* silencing in human leukemia cell lines (y-axis). The dotted lines represent the control condition (scramble transfected) 100%. Data is presented using the mean ± standard error of the mean. Significant differences from control conditions were indicated as * for *P*-value < 0.05, ** for *P*-value < 0.01, *** for *P*-value <0.001.

high and low risk patients independently of the cohort (Figure 5B) (Supplementary Figure S21). Furthermore, events that differentiate the high and low-risk patients, as well as the genes where they occur, had a small overlap with those that differentiate between KMT2A-r and ETV6-r patients (Supplementary Figure S22; Data file 9). This suggests that similar to the expression signature described above, there may be a splicing signature associated with risk that is independent of the fusion background.

To evaluate the SFs/RBPs potentially associated with these RNA processing changes, we performed a motif enrichment analysis. Specifically, we identified motifs for RBPs that were part of our high-risk signature: *SAMD4A* in A3 and AL events, *RBM24* in all the different types of events, *PCBP3* in AL and SE events, *PABPC4* in AL, and *MBNL1* in A3, AL, RI and SE events. We also recovered motifs for other RBPs, including the SR protein genes *SRSF1*, *SRSF2*, *SRSF7* and *SRSF9*. Interestingly, these

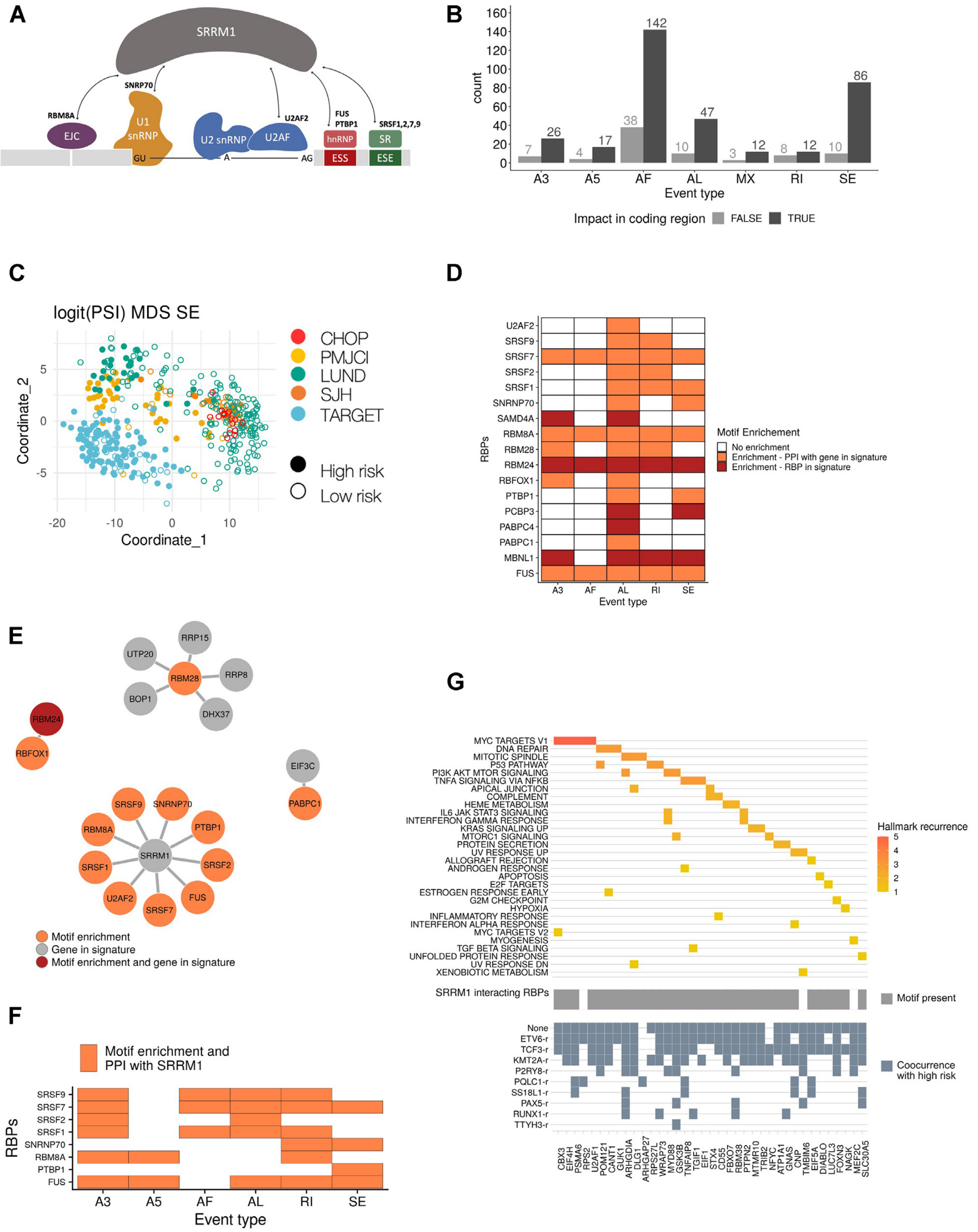


Figure 5. Splicing signature associated with high-risk. (**A**) SRRM1 is a known interactor of multiple splicing factors and RNA binding proteins (RBPs). (**B**) Identified significant splicing changes between high and low-risk patients. The counts are separated according to whether the splicing change impacts a coding region (black) or not (grey), and by event type: alternative 3′ (A3) and 5′ (A5) splice site, alternative first (AF), and alternative last (AL) exon, mutually exclusive exons (MX), retained intron (RI) and skipping exon (SE). (**C**) Multidimensional scaling (MDS) plot of the analyzed samples using only the significant SE events. The color indicates the patient cohort, and the full (empty) circle indicates if the sample is predicted to be of high (low) risk. Other event types are shown in Supplementary Figure S22. (**D**) Enrichment (z-score > 1.5) of binding motifs for RNA binding proteins (RBP) (y-axis) in each event type (x-axis). The plot indicates whether the RBP is part of the high-risk signature (red) or whether it has a reported protein-protein interaction (PPI) with a gene from the signature (orange). Event types not showing on the figure means that there are no binding motifs of RBP associated with the events. (**E**) Protein-protein interaction networks for the RBPs in the predictive signature and/or with binding motifs enriched in the events differentially spliced between high and low-risk patients. The color indicates whether the RBP motif was enriched (orange), whether the RBP was part of the high-risk signature (gray), or both (red). (**F**) RBPs and splicing factors with associated motifs enriched in events differentially spliced between mature B-cells and B-cell progenitors. (**G**) Upper panel shows cancer hallmarks associated with genes with differentially spliced events that co-occur with high risk in at least one fusion group, colored by its recurrence. In grey, the middle panel indicates the presence of binding motifs for any of the RBPs that interact with SRRM1. Lower panel highlights in blue the fusion groups in which the differentially spliced events co-occur with high risk.

and other SFs/RBPs with motifs enriched in events showed protein-protein interactions with our risk signature members (Figure 5D) (Supplementary Figure S23). We identified an enrichment in motifs for RBM28, an interactor of 5 of the proteins encoded by genes of our high-risk signature: *RRP8*, *RRP15*, *UTP20*, *BOP1*, and *DHX37* (Figure 5D). Similarly, we found an enrichment of motifs for *RBFOX1*, which interacts with *RBM24*, also part of the signature (Figure 5E). Remarkably, there were 9 RBPs with motif enrichment in the events changing with the risk that had protein-protein interactions with the splicing factor SRRM1 (Figure 5E). Moreover, SRRM1 expression correlated with the inclusion of events differentially spliced between high and low-risk patients, for events harboring binding motifs for RBPs with evidence of protein-protein interaction with SRRM1 (Supplementary Figure S24). Such correlation was not always observed for RBPs with enriched motifs and interacting with SRRM1. Moreover, SRRM1 was not among the splicing factors with the highest number of potential interactors (Supplementary Figure S25). These analyses suggest that a high-risk phenotype may be associated with a change in the expression of multiple SFs and RBPs that impact RNA processing, with a potentially prominent role for *SRRM1*.

We observed before that *SRRM1* expression as well as our risk signature score were higher in B-cells progenitors compared with mature B-cells. We thus next tested whether SRRM1 could play a role in the splicing changes across B-cell differentiation, we calculated the differentially spliced events and performed motif enrichment between progenitors and mature B-cells. Importantly, differentially spliced events between progenitors and mature B-cells contained motifs for the same SRRM1-interacting RBPs and splicing factors that we observed for leukemia patients (Figure 5F). Additionally, the genes with differentially spliced events were enriched in the same pathways obtained in the comparison between high and low risk patients, which were different from the pathways enriched in genes with splicing differences between *KMT2A-r* and *ETV6-r* (Supplementary Figure S26) (Data file 10). Interestingly, the expression level of the *KMT2A-AFF1* and *ETV6-RUNX1* fusions, the most abundantly observed in the *KMT2A-r* and *ETV6-r* groups, did not show any correlation with the risk score or with the *SRRM1* expression levels (Supplementary Figure S27).

To further investigate the possible mechanisms linking *SRRM1* with high risk, we calculated the subset of splicing events associated with risk that was also significantly associated with risk within each fusion group. Most of these events co-occurred with patients with *KMT2A-r*, *ETV6-r*, *TCF3-r*, as well as patients with no fusions (Figure 5G, lower panel) (Supplementary Figure S28); and presented binding motifs for RBPs interacting with *SRRM1* (Figure 5G, middle panel). Moreover, the genes harboring those events were significantly associated with cancer-related pathways, such as MYC targets, DNA repair and the p53 pathway (Figure 5g, upper panel). One of these genes was *EIF4H*, a translation initiation factor that is key for translational control. Overexpression of *EIF4H* has been associated before with cell proliferation and increased chemoresistance in lung cancer (57). Our analyses indicated that one of the EIF4H isoforms (ENST00000265753.12) decreased expression in the low-risk group, while a second isoform (ENST000000353999.6) had stable expression across all patients (Supplementary Figure S29).

Taken together, our results provide suggestive evidence that there is a molecular signature of expression and splicing changes, possibly driven by *SRRM1* that is predictive of bad prognosis in B-ALL and that is independent of the fusion background.

DISCUSSION

In this study, we performed a multicohort age-agnostic analysis of B-ALL cases focused on the differences of risk outcome independent of the fusion background. We identified an expression pattern involving MYC targets, translational regulators, and splicing factors associated with an increased probability of relapse. This is consistent with previous results showing that translation is tightly controlled during normal hematopoiesis (52) and is commonly deregulated in cancers, including hematological malignancies (58,59). Overexpression of *MYC* promotes expression of the translational machinery, increasing ribosome production and activity, leading to increased cell growth (53). Additionally, *MYC* overexpression plays a role in the dysregulation of the splicing machinery during lymphomagenesis (54). Our findings suggest that alterations in RNA processing could be involved in driving tumor progression and resistance to current therapies in B-ALL. This is consistent with recent work showing that aberrant splicing is directly implicated in the development of therapy resistance in B-ALL (60–63).

We summarized our findings in a 37-gene signature that showed prognostic value on B-ALL patients samples. Our signature classified patients in high and low risk with high accuracy and independently of the fusion background. The same signature separated high-risk patients from normal blood and spleen samples, and B-ALL cell lines and B-cell progenitors from mature B-cells. These results provide evidence that high-risk B-ALL cases recapitulate a gene expression pattern independently of the gene fusion background. In further support of our findings, analysis of an independent cohort (21) showed that our signature provided a significant separation of patients according to their relapse (log-rank test *P*-value 0.00041; Supplementary Figure S30a).

Our findings suggest that gene fusions operate mainly as initiating events, and an independent convergent mechanism defines high risk in B-ALL. This would agree with results using current murine and humanized models of *KMT2A-r* B-ALL (64) showing that they do not faithfully recapitulate the disease pathogenesis and suggesting that *KMT2A-r* alone is insufficient to sustain leukemia (65,66). Our derived gene expression signature could complement current clinical assessment methods of B-ALL patients. Our signature would be beneficial as a strategy to identify patients with a high risk of relapse when no fusions are detected, or when the presented fusion is of unknown prognosis. Moreover, it is conceivable that a deeper understanding of these genes may provide further mechanistic insight into the common functional underpinnings underlying B-ALL and thus reveal potential novel therapeutic avenues. Our high-risk signature included several splicing factors (SFs) and genes encoding for RNA binding proteins (RBPs), such as *SRRM1*, *IGF2BP2* and *MBNL1,* which suggested a pattern of differential RNA processing linked to high risk. MBNL1 is a protein involved in alternative splicing, predominantly regulates intron exclusion, and has been found consistently overexpressed in *KMT2A*-rearranged leukemia. Although inhibition of MBNL1 is linked with selective leukemic cell death, this effect seems to be *KMT2A*-r specific (67). Furthermore, the relative risk predictive power of *MBNL1* in our signature is low, possibly due to our cohorts having many patients with no *KMT2A* fusions, further supporting the specificity of MBNL1 to *KMT2A*-r cases.

Among the genes in our risk signature involved in RNA processing, *SRRM1* showed one of the strongest predictive powers and had the strongest correlation with the risk signature score. Samples at diagnosis from high-risk patients, before treatment, presented the highest *SRRM1* expression, similarly to samples obtained at relapse, and higher than normal blood/spleen samples. Moreover, *SRRM1* was highly expressed in B-ALL cell lines and B-cell progenitors compared to mature B-cells, further linking SRRM1 to a potential proliferative cellular state. Our analysis also showed that most of the splicing events changing between high and low-risk patients showed a correlation of their inclusion levels with SRRM1 expression and contained binding motifs for SFs/RBPs that have evidence of protein-protein interactions with SRRM1. Moreover, these events occurred in genes involved in cancer pathways. Importantly, analysis of an independent cohort (21) provided additional validation of the potential association of SRRM1 and its interactors with these splicing changes. Indeed, the splicing events changing between high and low-risk patients showed the same correlation patterns with SRRM1 and its RBP interactors in this new independent cohort (Supplementary Figure S30b).

Overexpression of *SRRM1* has been associated previously with poor prognosis in prostate cancer (55) and silencing of *SRRM1* was shown to reduce cell proliferation through a reduction of AKT phosphorylation levels and an increased expression of *PTEN*, a well-known tumor suppressor (55). Previous studies have also shown that *SRRM1* overexpression leads to the expression of a *CD44* isoform that acts as a RAS-signaling activator and induces metastatic potential in non-metastatic cells (68). Furthermore, *SRRM1* has been identified as part of a chromatin protein complex that drives B-cell differentiation (69). We showed that silencing of *SRRM1* in B-ALL cell models leads to a significant decrease in proliferation and invasion rates and a significant increase in apoptosis capacity. The effect in response to *SRRM1* silencing was different in all cell lines tested, especially in the decrease in proliferation rate. This appeared to be associated with the variable *SRRM1* expression found in these cell lines rather than with the basal proliferation rate of each cell model. This would indicate that tumors with higher *SRRM1* expression would depend more on *SRRM1* for proliferation. In contrast, tumors with lower *SRRM1* expression would rely on other proliferation mechanisms. Importantly, we analysed data from a recent pan-cancer protein map atlas based on 946 human cancer cell lines (70), and found that SRRM1 presents the highest protein levels in hematological tumors (Supplementary Figure S31). This provides an additional layer of evidence for the potential role of SRRM1 in the progression of leukemia and, in particular, B-ALL.

In conclusion, we have presented a gene expression signature that predicts poor outcomes in samples at diagnosis independently of the fusion background. This signature is associated with *SRRM1* overexpression and with splicing changes potentially partly driven by SRRM1 interactions with other splicing factors. This leads us to propose that *SRRM1* overexpression may contribute to sustaining tumor malignancy and lead to poor prognosis in B-ALL. Furthermore, *SRRM1* could function as a novel prognostic marker of high-risk B-ALL, and its depletion could be used in combination with standard therapies to achieve more effective treatments in high-risk cases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Cancer Online.

ACKNOWLEDGEMENTS

We thank all the authors involved in the different studies used in this article for facilitating access to their datasets. We thank Alfons Gonzalez for his continued technical support. The funding bodies did not contribute to the design of the study, analysis or interpretation of data, nor to the writing of the manuscript.

FUNDING

Spanish Ministerio de Ciencia, Innovación y Universidades [BIO2017-85364-R]; EMBL Australia (to A.C., M.R-S., J.L.M., E.E.); National Institutes of Health [U01 CA232563 to A.T-T.]; M.T-D. acknowledges support from the Ellen Weisberg Fund: Advancing Breakthroughs in Pediatric Cancer. This work was further supported by the Spanish Ministry of Science and Innovation [MICINN, PID2019-105201RB-I00 to J.P.C.]; Junta de Andalucía [BIO-0139]; Universidad de Córdoba-FEDER [UCO-202099901918904 to J.P.C.]; GETNE2019 Research grant (to J.P.C.); CIBERobn Fisiopatología de la Obesidad y Nutrición (CIBER is an initiative of Instituto de Salud Carlos III, co-funded by the European Union: ERDF/ESF, 'Investing in your future').

Conflict of interest statement. None declared.

REFERENCES

1. Pui,C.-H. and Evans,W.E. (2013) A 50-year journey to cure childhood acute lymphoblastic leukemia. *Semin. Hematol.*, **50**, 185–196.
2. Gröbner,S.N., Worst,B.C., Weischenfeldt,J., Buchhalter,I., Kleinheinz,K., Rudneva,V.A., Johann,P.D., Balasubramanian,G.P., Segura-Wang,M., Brabetz,S. *et al.* (2018) the landscape of genomic alterations across childhood cancers. *Nature*, **555**, 321–327.
3. Boer,J.M. and Boer,M.L. (2017) BCR-ABL1-like acute lymphoblastic leukaemia: from bench to bedside. *Eur. J. Cancer*, **82**, 203–218.
4. Fischer,U., Forster,M., Rinaldi,A., Risch,T., Sungalee,S., Warnatz,H.-J., Bornhauser,B., Gombert,M., Kratsch,C., Stütz,A.M. *et al.* (2015) Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options. *Nat. Genet.*, **47**, 1020–1029.

5. Pikman,Y. and Stegmaier,K. (2018) Targeted therapy for fusion-driven high-risk acute leukemia. *Blood*, **132**, 1241–1247.
6. Wang,Y., Wu,N., Liu,D. and Jin,Y. (2017) Recurrent fusion genes in leukemia: an attractive target for diagnosis and treatment. *Curr. Genomics*, **18**, 378–384.
7. Bueno,C., Montes,R., Catalina,P., Rodríguez,R. and Menendez,P. (2011) Insights into the cellular origin and etiology of the infant pro-B acute lymphoblastic leukemia with MLL-AF4 rearrangement. *Leukemia*, **25**, 400–410.
8. Meyer,C., Burmeister,T., Gröger,D., Tsaur,G., Fechina,L., Renneville,A., Sutton,R., Venn,N.C., Emerenciano,M., Pombo-de-Oliveira,M.S. *et al.* (2018) the MLL recombinome of acute leukemias in 2017. *Leukemia*, **32**, 273–284.
9. Montes,R., Ayllón,V., Gutierrez-Aranda,I., Prat,I., Hernández-Lamas,M.C., Ponce,L., Bresolin,S., Te Kronnie,G., Greaves,M., Bueno,C. *et al.* (2011) Enforced expression of MLL-AF4 fusion in cord blood CD34+ cells enhances the hematopoietic repopulating cell function and clonogenic potential but is not sufficient to initiate leukemia. *Blood*, **117**, 4746–4758.
10. Bursen,A., Schwabe,K., Rüster,B., Henschler,R., Ruthardt,M., Dingermann,T. and Marschalek,R. (2010) the AF4·MLL fusion protein is capable of inducing ALL in mice without requirement of MLL·AF4. *Blood*, **115**, 3570–3579.
11. Bueno,C., Montes,R., Melen,G.J., Ramos-Mejia,V., Real,P.J., Ayllón,V., Sanchez,L., Ligero,G., Gutierrez-Aranda,I., Fernández,A.F. *et al.* (2012) a human ESC model for MLL-AF4 leukemic fusion gene reveals an impaired early hematopoietic-endothelial specification. *Cell Res.*, **22**, 986–1002.
12. Krivtsov,A.V and Armstrong,S.A. (2007) MLL translocations, histone modifications and leukaemia stem-cell development. *Nat. Rev. Cancer*, **7**, 823–833.
13. Tasian,S.K., Loh,M.L. and Hunger,S.P. (2017) Philadelphia chromosome-like acute lymphoblastic leukemia. *Blood*, **130**, 2064–2072.
14. Andersson,A.K., Ma,J., Wang,J., Chen,X., Gedman,A.L., Dang,J., Nakitandwe,J., Holmfeldt,L., Parker,M., Easton,J. *et al.* (2015) The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nat. Genet.*, **47**, 330–337.
15. Lilljebjörn,H., Henningsson,R., Hyrenius-Wittsten,A., Olsson,L., Orsmark-Pietras,C., von Palffy,S., Askmyr,M., Rissler,M., Schrappe,M., Cario,G. *et al.* (2016) Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat. Commun.*, **7**, 11790.
16. Black,K.L., Naqvi,A.S., Asnani,M., Hayer,K.E., Yang,S.Y., Gillespie,E., Bagashev,A., Pillai,V., Tasian,S.K., Gazzara,M.R. *et al.* (2018) Aberrant splicing in B-cell acute lymphoblastic leukemia. *Nucleic Acids Res.*, **46**, 11357–11369.
17. Agraz-Doblas,A., Bueno,C., Bashford-Rogers,R., Roy,A., Schneider,P., Bardini,M., Ballerini,P., Cazzaniga,G., Moreno,T., Revilla,C. *et al.* (2019) Unraveling the cellular origin and clinical prognostic markers of infant B-cell acute lymphoblastic leukemia using genome-wide analysis. *Haematologica*, **104**, 1176–1188.
18. Ghandi,M., Huang,F.W., Jané-Valbuena,J., Kryukov,G.V, Lo,C.C., McDonald,E.R., Barretina,J., Gelfand,E.T., Bielski,C.M., Li,H. *et al.* (2019) Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, **569**, 503–508.
19. Carithers,L.J., Ardlie,K., Barcus,M., Branton,P.A., Britton,A., Buia,S.A., Compton,C.C., DeLuca,D.S., Peter-Demchok,J., Gelfand,E.T. *et al.* (2015) a Novel Approach to High-Quality Postmortem Tissue Procurement: the GTEx Project. *Biopreserv. Biobank.*, **13**, 311–319.
20. Yang,S.Y., Hayer,K.E., Fazelinia,H., Spruce,L.A., Asnani,M., Black,K.L., Naqvi,A.S., Pillai,V., Barash,Y., Elenitoba-Johnson,K.S.J. *et al.* (2022) FBXW7β isoform drives transcriptional activation of the proinflammatory TNF cluster in human pro-B cells. *Blood Adv.*, https://doi.org/10.1182/bloodadvances.2022007910.
21. Jeha,S., Choi,J., Roberts,K.G., Pei,D., Coustan-Smith,E., Inaba,H., Rubnitz,J.E., Ribeiro,R.C., Gruber,T.A., Raimondi,S.C. *et al.* (2021) Clinical significance of novel subtypes of acute lymphoblastic leukemia in the context of minimal residual disease-directed therapy. *Blood Cancer Discov.*, **2**, 326–337.
22. Andrews,S. (2010) FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
23. Haas,B.J., Dobin,A., Li,B., Stransky,N., Pochet,N. and Regev,A. (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.*, **20**, 213.
24. Hsieh,G., Bierman,R., Szabo,L., Lee,A.G., Freeman,D.E., Watson,N., Sweet-Cordero,E.A. and Salzman,J. (2017) Statistical algorithms improve accuracy of gene fusion detection. *Nucleic Acids Res.*, **45**, e126.
25. Babiceanu,M., Qin,F., Xie,Z., Jia,Y., Lopez,K., Janus,N., Facemire,L., Kumar,S., Pang,Y., Qi,Y. *et al.* (2016) Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Res.*, **44**, 2859–2872.
26. Gao,Q., Liang,W.-W., Foltz,S.M., Mutharasu,G., Jayasinghe,R.G., Cao,S., Liao,W.-W., Reynolds,S.M., Wyczalkowski,M.A., Yao,L. *et al.* (2018) Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.*, **23**, 227–238.
27. Quek,L., Otto,G.W., Garnett,C., Lhermitte,L., Karamitros,D., Stoilova,B., Lau,I.-J., Doondeea,J., Usukhbayar,B., Kennedy,A. *et al.* (2016) Genetically distinct leukemic stem cells in human CD34- acute myeloid leukemia are arrested at a hemopoietic precursor-like stage. *J. Exp. Med.*, **213**, 1513–1535.
28. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2019) COSMIC: the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **47**, D941–D947.
29. Mitelman,F., Johansson,B. and Mertens,F. (2020) Mitelman database of chromosome aberrations and gene fusions in cancer. https://mitelmandatabase.isb-cgc.org.
30. Constantinescu,S., Szczurek,E., Mohammadi,P., Rahnenführer,J. and Beerenwinkel,N. (2016) TiMEx: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics*, **32**, 968–975.
31. Smedley,D., Haider,S., Durinck,S., Pandini,L., Provero,P., Allen,J., Arnaiz,O., Awedh,M.H., Baldock,R., Barbiera,G. *et al.* (2015) the BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.*, **43**, W589–W598.
32. Frankish,A., Diekhans,M., Ferreira,A.-M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.
33. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
34. Soneson,C., Love,M.I. and Robinson,M.D. (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, **4**, 1521.
35. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
36. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
37. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.
38. Sebestyén,E., Singh,B., Miñana,B., Pagès,A., Mateo,F., Pujana,M.A., Valcárcel,J. and Eyras,E. (2016) Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.*, **26**, 732–744.
39. Singh,B., Trincado,J.L., Tatlow,P.J., Piccolo,S.R. and Eyras,E. (2018) Genome sequencing and RNA-motif analysis reveal novel damaging noncoding mutations in human tumors. *Mol. Cancer Res.*, **16**, 1112–1124.
40. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
41. Alamancos,G.P., Pagés,A., Trincado,J.L., Bellora,N. and Eyras,E. (2015) Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*, **21**, 1521–1531.

42. Trincado,J.L., Entizne,J.C., Hysenaj,G., Singh,B., Skalic,M., Elliott,D.J. and Eyras,E. (2018) SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.*, **19**, 40..
43. Hastie,T., Tibshirani,R., Narasimhan,B. and Chu,G. (2020) impute: impute: Imputation for microarray data. https://bioconductor.org/packages/release/bioc/html/impute.html.
44. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
45. Griffith,D.M., Veech,J.A. and Marsh,C.J. (2016) cooccur: probabilistic species co-occurrence analysis in R. *J. Stat. Softw.*, **69**, 1–17.
46. Yu,G., Wang,L.-G., Han,Y. and He,Q.-Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, **16**, 284–287.
47. Uphoff,C.C. and Drexler,H.G. (2013) Detection of mycoplasma contaminations. *Methods Mol. Biol.*, **946**, 1–13.
48. Luque,R.M., Ibáñez-Costa,A., Neto,L.V., Taboada,G.F., Hormaechea-Agulla,D., Kasuki,L., Venegas-Moreno,E., Moreno-Carazo,A., Gálvez,M.Á., Soto-Moreno,A. *et al.* (2015) Truncated somatostatin receptor variant sst5TMD4 confers aggressive features (proliferation, invasion and reduced octreotide response) to somatotropinomas. *Cancer Lett.*, **359**, 299–306.
49. Fuentes-Fayos,A.C., Vázquez-Borrego,M.C., Jiménez-Vacas,J.M., Bejarano,L., Pedraza-Arévalo,S., L.-López,F., Blanco-Acevedo,C., Sánchez-Sánchez,R., Reyes,O., Ventura,S. *et al.* (2020) Splicing machinery dysregulation drives glioblastoma development/aggressiveness: oncogenic role of SRSF3. *Brain*, **143**, 3273–3293.
50. Ma,X., Edmonson,M., Yergeau,D., Muzny,D.M., Hampton,O.A., Rusch,M., Song,G., Easton,J., Harvey,R.C., Wheeler,D.A. *et al.* (2015) Rise and fall of subclones from diagnosis to relapse in pediatric B-acute lymphoblastic leukaemia. *Nat. Commun.*, **6**, 6604.
51. Ning,J., Zhao,Y., Ye,Y. and Yu,J. (2019) Opposing roles and potential antagonistic mechanism between TGF-β and BMP pathways: implications for cancer progression. *EBioMedicine*, **41**, 702–710.
52. Signer,R.A.J., Magee,J.A., Salic,A. and Morrison,S.J. (2014) Haematopoietic stem cells require a highly regulated protein synthesis rate. *Nature*, **509**, 49–54.
53. Hodson,D.J., Screen,M. and Turner,M. (2019) RNA-binding proteins in hematopoiesis and hematological malignancy. *Blood*, **133**, 2365–2373.
54. Koh,C.M., Bezzi,M., Low,D.H.P., Ang,W.X., Teo,S.X., Gay,F.P.H., Al-Haddawi,M., Tan,S.Y., Osato,M., Sabò,A. *et al.* (2015) MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. *Nature*, **523**, 96–100.
55. Jiménez-Vacas,J.M., Herrero-Aguayo,V., Montero-Hidalgo,A.J., Gómez-Gómez,E., Fuentes-Fayos,A.C., León-González,A.J., Sáez-Martínez,P., Alors-Pérez,E., Pedraza-Arévalo,S., González-Serrano,T. *et al.* (2020) Dysregulation of the splicing machinery is directly associated to aggressiveness of prostate cancer. *EBioMedicine*, **51**, 102547.
56. Blencowe,B.J., Issner,R., Nickerson,J.A. and Sharp,P.A. (1998) a coactivator of pre-mRNA splicing. *Genes Dev.*, **12**, 996–1009.
57. Vaysse,C., Philippe,C., Martineau,Y., Quelen,C., Hieblot,C., Renaud,C., Nicaise,Y., Desquesnes,A., Pannese,M., Filleron,T. *et al.* (2015) Key contribution of eIF4H-mediated translational control in tumor promotion. *Oncotarget*, **6**, 39924–39940.
58. Silvera,D., Formenti,S.C. and Schneider,R.J. (2010) Translational control in cancer. *Nat. Rev. Cancer*, **10**, 254–266.
59. Delgado,M.D. and Leon,J. (2010) Myc roles in hematopoiesis and leukemia. *Genes Cancer*, **1**, 605–616.
60. Zheng,S., Gillespie,E., Naqvi,A.S., Hayer,K.E., Ang,Z., Torres-Diz,M., Quesnel-Vallières,M., Hottman,D.A., Bagashev,A., Chukinas,J. *et al.* (2022) Modulation of CD22 protein expression in childhood leukemia by pervasive splicing aberrations: implications for CD22-directed immunotherapies. *Blood Cancer Discov.*, **3**, 103–115.
61. Sotillo,E., Barrett,D.M., Black,K.L., Bagashev,A., Oldridge,D., Wu,G., Sussman,R., Lanauze,C., Ruella,M., Gazzara,M.R. *et al.* (2015) Convergence of acquired mutations and alternative splicing of CD19 enables resistance to CART-19 immunotherapy. *Cancer Discov.*, **5**, 1282–1295.
62. Asnani,M., Hayer,K.E., Naqvi,A.S., Zheng,S., Yang,S.Y., Oldridge,D., Ibrahim,F., Maragkakis,M., Gazzara,M.R., Black,K.L. *et al.* (2020) Retention of CD19 intron 2 contributes to CART-19 resistance in leukemias with subclonal frameshift mutations in CD19. *Leukemia*, **34**, 1202–1207.
63. Cortés-López,M., Schulz,L., Enculescu,M., Paret,C., Spiekermann,B., Quesnel-Vallières,M., Torres-Diz,M., Unic,S., Busch,A., Orekhova,A. *et al.* (2022) High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling CD19 splicing and CART-19 therapy resistance. *Nat. Commun.*, **13**, 5570.
64. Lin,S., Luo,R.T., Shrestha,M., Thirman,M.J. and Mulloy,J.C. (2017) the full transforming capacity of MLL-Af4 is interlinked with lymphoid lineage commitment. *Blood*, **130**, 903–907.
65. Prieto,C., Stam,R.W., Agraz-Doblas,A., Ballerini,P., Camos,M., Castaño,J., Marschalek,R., Bursen,A., Varela,I., Bueno,C. *et al.* (2016) Activated KRAS cooperates with MLL-AF4 to promote extramedullary engraftment and migration of cord blood CD34+ HSPC but is insufficient to initiate leukemia. *Cancer Res.*, **76**, 2478–2489.
66. Bueno,C., Ayllón,V., Montes,R., Navarro-Montero,O., Ramos-Mejia,V., Real,P.J., Romero-Moya,D., Araúzo-Bravo,M.J. and Menendez,P. (2013) FLT3 activation cooperates with MLL-AF4 fusion protein to abrogate the hematopoietic specification of human ESCs. *Blood*, **121**, 3867–3878.
67. Itskovich,S.S., Gurunathan,A., Clark,J., Burwinkel,M., Wunderlich,M., Berger,M.R., Kulkarni,A., Chetal,K., Venkatasubramanian,M., Salomonis,N. *et al.* (2020) MBNL1 regulates essential alternative RNA splicing patterns in MLL-rearranged leukemia. *Nat. Commun.*, **11**, 2369.
68. Cheng,C. and Sharp,P.A. (2006) Regulation of CD44 alternative splicing by SRm160 and its potential role in tumor cell invasion. *Mol. Cell. Biol.*, **26**, 362–370.
69. Ochiai,K., Yamaoka,M., Swaminathan,A., Shima,H., Hiura,H., Matsumoto,M., Kurotaki,D., Nakabayashi,J., Funayama,R., Nakayama,K. *et al.* (2020) Chromatin protein PC4 orchestrates B cell differentiation by collaborating with IKAROS and IRF4. *Cell Rep.*, **33**, 108517.
70. Gonçalves,E., Poulos,R.C., Cai,Z., Barthorpe,S., Manda,S.S., Lucas,N., Beck,A., Bucio-Noble,D., Dausmann,M., Hall,C. *et al.* (2022) Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell*, **40**, 835–849.