# Aberrant splicing in B-cell acute lymphoblastic leukemia

Kathryn L. Black*[1], Ammar S. Naqvi*[1,2], Katharina E. Hayer[1,2], Scarlett Y. Yang[1,8], Elisabeth Gillespie[1#], Asen Bagashev[1], Vinodh Pillai[1], Sarah K. Tasian[3], Matthew R. Gazzara[4,5], Martin Carroll[6], Deanne Taylor[2,3], Kristen W. Lynch[4,8], Yoseph Barash[5], and Andrei Thomas-Tikhonenko[1,3,7,8]

Departments of [1] Pathology (Cancer Pathobiology), [2] Biomedical & Health Informatics (Bioinformatics) and [3] Pediatrics (Oncology and Center for Childhood Cancer Research), Children's Hospital of Philadelphia; Departments of [4] Biochemistry & Biophysics, [5] Genetics, [6] Medicine, and [7] Pathology & Laboratory Medicine, Perelman School of Medicine and [8] Immunology Graduate Group, University of Pennsylvania, Philadelphia, PA 19104.

*- equal contribution
# - present affiliation: Inovio Pharmaceuticals, Plymouth Meeting, PA 19462

***Corresponding author:***
*Andrei Thomas-Tikhonenko, Ph.D.*
*The Children's Hospital of Philadelphia*
*4056 Colket Translational Research Bldg*
*3501 Civic Center Blvd*
*Philadelphia, PA 19104-4399*
*Phone: 267-426-9699*
*Fax: 267-426-8125*
*E-mail: andreit@mail.med.upenn.edu*

## ABSTRACT

**Background:** Mutations in splicing factor (SF) genes cause myelodysplastic syndromes and chronic lymphocytic leukemia (CLL). While such mutations had not been found in B-cell acute lymphoblastic leukemia (B-ALL), we previously reported that alternative splicing of the CD19 transcript is a robust mechanism of resistance to CD19-directed immunotherapy in children with B-ALL. We thus hypothesized that additional mRNAs may be alternatively spliced in these leukemias.

**Results:** Using flow cytometry-based cell purification protocols, deep RNA sequencing (RNA-seq), and the MAJIQ algorithm, we compared transcriptomes of CD19+/CD34+ pro-B cells from normal bone marrow donors to 18 primary pediatric B-ALL samples. We found 4,000-5,000 differential local splice variations (LSV) per leukemia sample, with 279 LSVs in 241 genes differentially spliced in every B-ALL sample. The consistently mis-spliced genes were significantly enriched in the RNA splicing pathway components and encoded ~100 different SFs, many from the SRSF and hnRNP families. Since aberrant LSVs in hnRNPA1 mRNA were present in 100% of B-ALL samples, we knocked down this transcript in a B-lymphoblastoid cell line using siRNA and defined 213 robust hnRNPA1-dependent events. Nearly 30% of the hnRNPA1-dependent LSVs were detectable in B-ALL samples, with one of the affected genes being DICER1, which is commonly mutated or down-regulated in many hematologic malignancies and included in the COSMIC dataset. We next asked how many other COSMIC genes are affected by aberrant splicing. We discovered 81 LSVs (mainly hnRNPA1-independent) in 41 COSMIC genes, including FBXW7, which was alternatively spliced in all 18 primary B-ALL samples. We were able to confirm 77 out of 81 of these LSVs in at least one of the two large independent RNA-seq B-ALL datasets generated by the TARGET Consortium and St Jude Children's Research Hospital. In fact, the twenty most common B-ALL drivers showed much higher prevalence of aberrant splicing than of somatic mutations.

**Conclusions:** B-ALL has widespread changes in splicing, likely due to the aberrant exon usage by SF-encoding transcripts. Aberrant splicing also affects most known B-ALL drivers, suggesting that this type of post-transcriptional regulation contributes to disease pathogenesis.


**Keywords:** B-cell acute lymphoblastic leukemia, alternative splicing, RNA-Seq, bone marrow differentiation

## BACKGROUND

Despite advances in the treatment of pediatric B-ALL, children with relapsed or refractory disease account for a substantial number of childhood cancer-related deaths. Adults with B-ALL experience even higher relapse rates and long-term event-free survival of less than 50% [1]. Recently, significant gains in the treatment of B-ALL have been achieved through the use of immunotherapies directed against CD19, a protein expressed on the surface of most B-cell neoplasms [2, 3]. These gains culminated in the recent FDA approval of tisagenlecleucel and axicabtagene ciloleucel, CD19-redirected chimeric antigen receptor (CAR) T-cell immunotherapies, for patients with refractory/relapsed B-cell malignancies. However, relapses occur in 10-20% of patients with B-ALL treated with CD19 CAR T cells, often due to epitope loss and/or B-cell de-differentiation into other lineages [4-7]. Other targets for immunotherapy include CD20 and CD22 [8-11]. However, neither antigen is uniformly expressed in B-ALL, and factors accounting for this mosaicism are poorly understood [3]. Elucidating mechanisms controlling epitope presentation will be key to anticipating and counteracting resistance.

We previously reported new mechanisms of pediatric B-ALL resistance to CD19-targeting immunotherapy. We discovered that in some cases, resistance to CD19 CAR T cells was generated through alternative splicing of CD19 transcripts. This post-transcriptional event was mediated by a specific splicing factor (SF) SRSF3 and generated a CD19 protein isoform invisible to the immunotherapeutic agent through skipping of exon 2 ([12], reviewed by [13, 14]).

Our discovery of a resistance mechanism based on alternative splicing prompted us to investigate the extent of this phenomenon in additional B-ALL cases. While driver mutations in splicing factors such as SRSF2, SF3B1, and U2AF1 have recently been discovered in myelodysplastic syndrome/acute myelogenous leukemia [15-17] and chronic lymphocytic leukemia [18, 19], SF mutations have not been reported in B-ALL. Nevertheless, our prior work suggested the possibility that SRSF3 (and by inference other SFs) could be deregulated in B-ALL [12], bringing about wide-spread splicing aberration.

This model would be particularly attractive because B-ALL is a heterogeneous, chromosome translocation-driven disease where the prevalence of somatic mutations and copy number variations is relatively low. For example, the commonly mutated *IKZF1* gene (which encodes the Ikaros transcription factor) is affected by missense mutations in just ~20% of B-ALL cases. Similarly, mutations in the key tumor suppressor gene (TSG) TP53 are found in only ~7% of B-ALLs (data from the COSMIC database) [20, 21]. In addition, both genes are robustly transcribed across individual B-ALLs and thus are not epigenetically silenced. This raises the possibility that they and other TSGs are dysregulated by post-transcriptional events, such as alternative splicing.

## RESULTS

### RNA-seq analysis of bone marrow-derived human B-cells.

To determine if patterns of splicing dysregulation occur in B-ALL, we first generated normal B-lymphocyte datasets corresponding to potential cells of origin. To this end, we obtained from the University of Pennsylvania Stem Cell and Xenograft Core facility two healthy adult bone marrow biopsies and from the Children's Hospital of Philadelphia (CHOP) Hematopathology Laboratory two bone marrow biopsies from children without leukemia. We enriched for mononuclear cells using Ficoll gradient separation, stained cells for combinations of stage-specific surface markers, and sorted B-cell subsets using flow cytometry. Specifically, we fractionated bone marrow progenitors into early progenitors ($CD34^+/CD19^-/IgM^-$), pro-B ($CD34^+/CD19^+/IgM^-$), pre-B ($CD34^-/CD19^+/IgM^-$), and immature B ($CD34^-/CD19^+/IgM^+$) populations (Fig. 1a). We then extracted RNA, performed RNA-seq, and quantified transcript levels. Concordant with flow cytometric profiles, CD19 mRNA levels increased

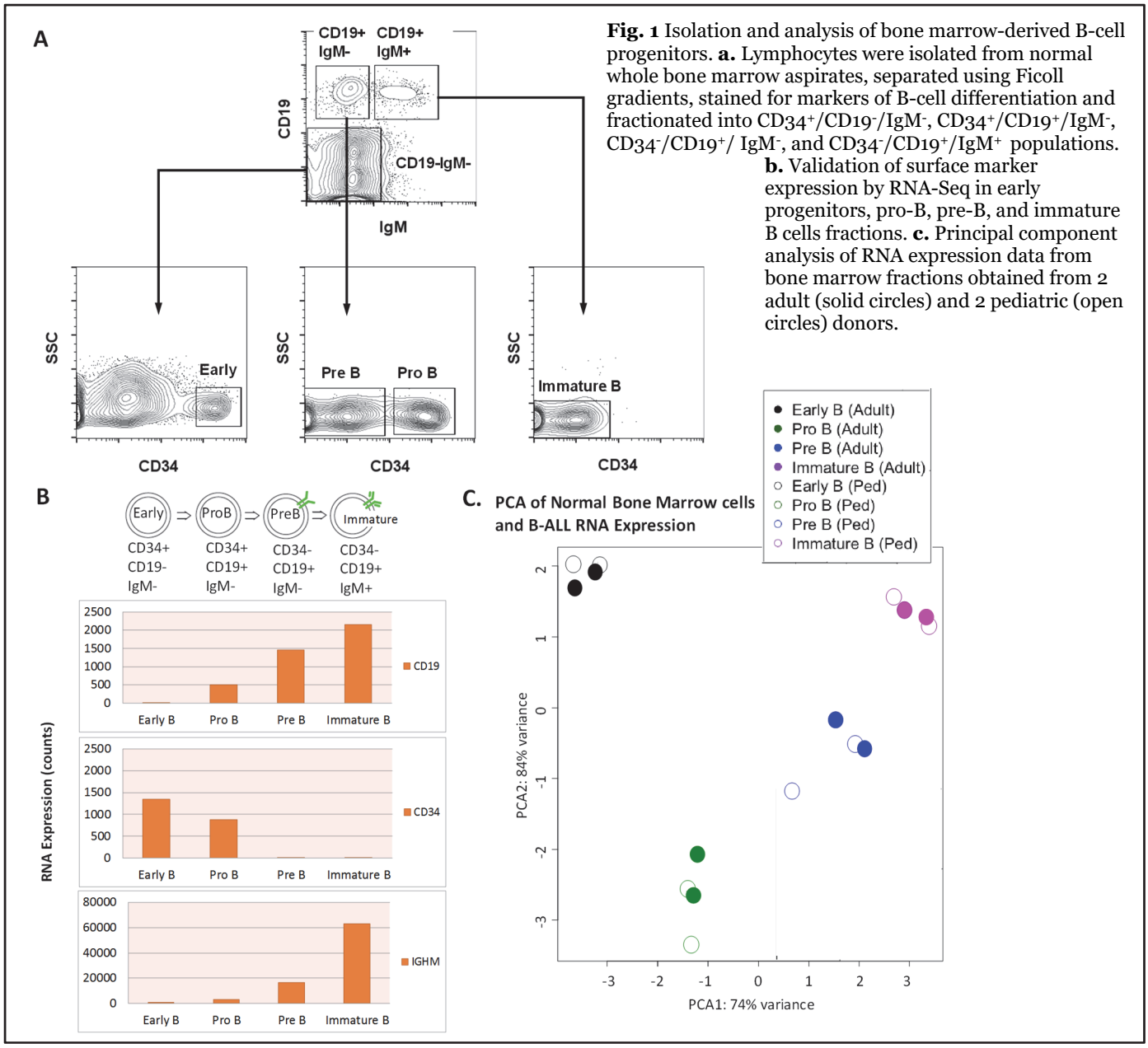


**Fig. 1** Isolation and analysis of bone marrow-derived B-cell progenitors. **a.** Lymphocytes were isolated from normal whole bone marrow aspirates, separated using Ficoll gradients, stained for markers of B-cell differentiation and fractionated into $CD34^+/CD19^-/IgM^-$, $CD34^+/CD19^+/IgM^-$, $CD34^-/CD19^+/IgM^-$, and $CD34^-/CD19^+/IgM^+$ populations. **b.** Validation of surface marker expression by RNA-Seq in early progenitors, pro-B, pre-B, and immature B cells fractions. **c.** Principal component analysis of RNA expression data from bone marrow fractions obtained from 2 adult (solid circles) and 2 pediatric (open circles) donors.

throughout differentiation stages, CD34 mRNA was confined to early progenitors and pro-B fractions, and IgM transcript was expressed only in the immature fraction (Fig. 1b). Furthermore, we performed principal component analysis (PCA) on all expression datasets and found tight clustering of the four fractions from different donors, suggesting that at the level of mRNA expression B-cell differentiation supersedes individual variations (Fig. 1c).

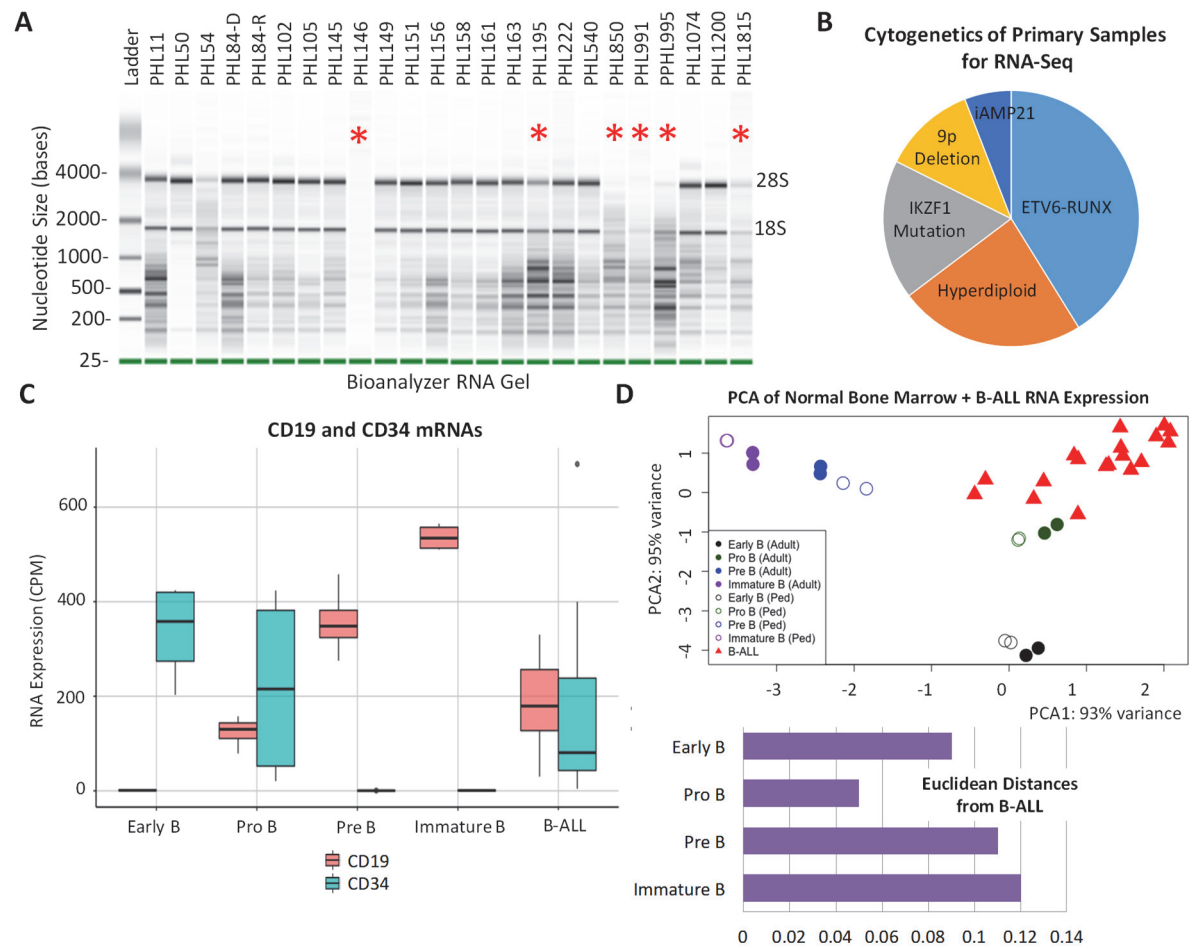


**Fig. 2** RNA analysis of pediatric B-ALL. **a.** RNA integrity (RIN) analysis of 24 cryopreserved leukemia samples. 18 of them passing QC criteria were chosen for Next Generation sequencing. Discarded samples are denoted with red asterisks. **b.** Pie chart showing distribution of cytogenetic abnormalities in 18 B-ALL samples. **c.** Quantification of CD19 and CD34 expression in bone marrow fractions and B-ALL samples. **d.** PCA on RNA expression from 18 B-ALL samples and 4 bone marrow samples. **e.** Euclidean distances showing closest clustering of B-ALL with pro-B cell fractions.

## RNA-Seq analysis of primary B-ALL samples

We obtained 24 primary pediatric B-ALL samples from the CHOP Center for Childhood Cancer Research leukemia biorepository. Mononuclear cells from fresh bone marrow or peripheral blood specimens were purified via Ficoll gradient separation (Fig. 1a) and cryopreserved for downstream experimental use. We extracted total RNA and analyzed sample integrity. Based on the presence of intact RNA (evident by 28S and 18S bands on gels), an RNA integrity score >5.2, and RNA concentration of >40ng/uL, we successfully extracted high quality RNA for sequencing from 18 out of 24 frozen samples (Fig. 2a, red asterisks indicate samples that did not pass QC and were not sequenced). These samples were comprised of several different phenotypic and genetic B-ALL subtypes [22] (Table 1 and Fig. 2b). We performed RNA-seq of these leukemia samples and first compared them

to adult and pediatric normal bone marrow fractions with respect to CD19 and CD34 expression. We found that B-ALL samples closely resembled the pro-B fractions in that CD19 and CD34 transcripts were readily detectable in the low-to-medium range (Fig. 2c). To extend this analysis to the entire transcriptome, we then performed PCA on B-ALL samples and four sets of normal bone marrow fractions (two pediatric and two adult). Once again, the 18 B-ALL samples clustered most closely with the pro-B fractions (Fig. 2d). This similarity was reflected in shortest Euclidian distances between PCA1 coordinates for the centroids of the B-ALL and pro-B cell samples (Fig. 2e). Therefore, we chose the pro-B fractions as cell-of-origin controls for B-ALL splicing analysis.

## Global patterns of aberrant splicing in pediatric B-ALL

To detect patterns of alternative splicing in B-ALL, we utilized the MAJIQ (Modeling Alternative Junction Inclusion Quantification) algorithm [23]. Using MAJIQ, we performed 18 independent pairwise comparisons between the average of pediatric pro-B cell fractions and leukemia samples. To assess heterogeneity in splicing across samples, we measured the number of differential LSVs (minimum of 20% change in splicing, 95% CI) in each sample and compared their identities. We observed that each B-ALL specimen had >3000 LSVs but most LSVs were either unique (first bar in Fig. 3a) or shared by only a small number of B-ALL patient samples (subsequent bars in Fig. 3a).

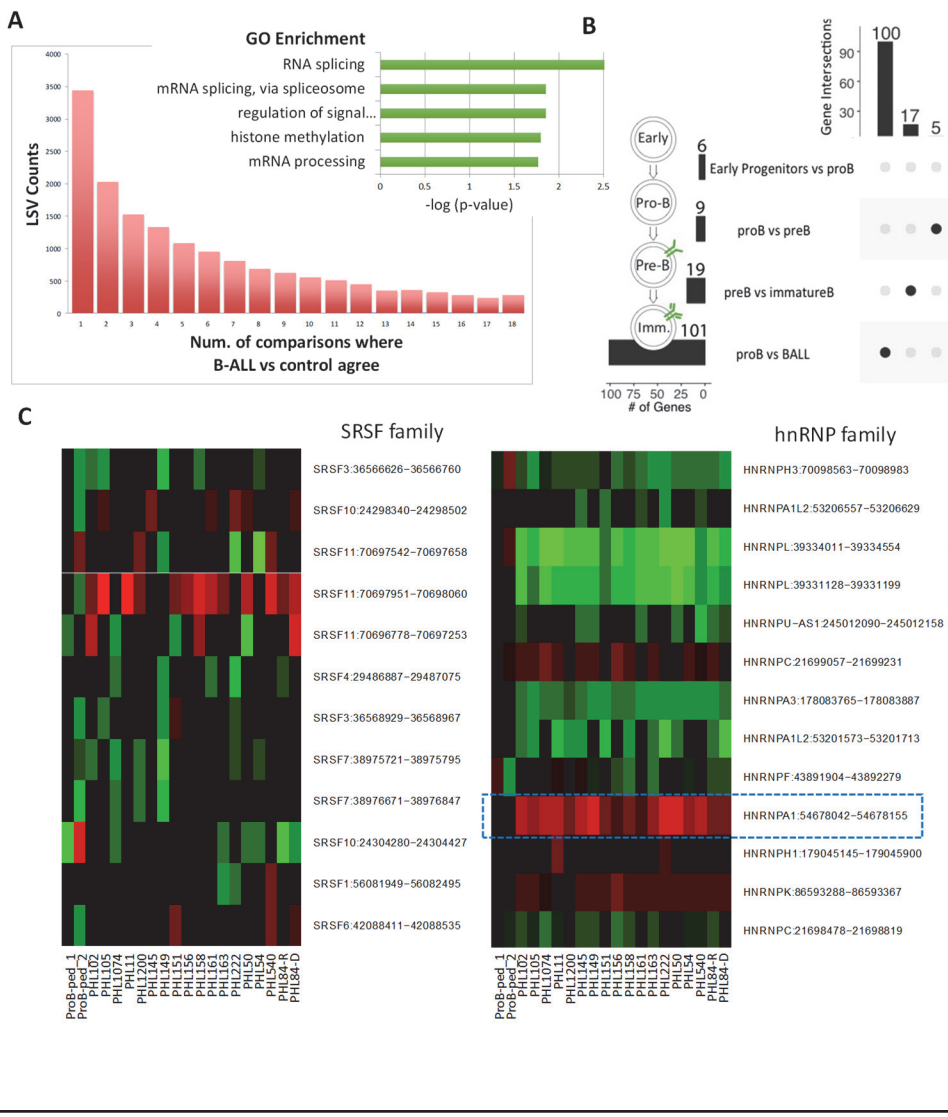


**Fig. 3.** Alternative splicing in pediatric B-ALL. **a.** The vertical bar chart shows the numbers of LSVs detected by MAJIQ per B-ALL sample compared to normal pro-B cells. The horizontal bar chart (insert) showing enrichment for RNA splicing pathways among genes with most consistent LSVs. **b**. Overlap of LSVs in SF genes between B-ALL samples and bone marrow fractions. The left panel shows the different sets and number of genes included in them. The middle section visualizes the different combinations of intersection (including each set by itself). The upper panel shows how many genes of the total set are in each intersection. **c.** Heat maps showing changes in inclusion of exons corresponding to SRSF and hnRNP families members in 18 B-ALL samples compared to the pro-B cell fraction. Green color denotes increased inclusion, red color – increased skipping.

However, we also found a total of 279 aberrant LSVs in 241 genes that were detected in all of our 18 pairwise comparisons (last bar in Fig. 3a). Provocatively, when this 241-gene set was analyzed using DAVID (Database for Annotation, Visualization, and Integrated Discovery) [24], the top gene ontology (GO) categories [25] were related to RNA splicing (Fig. 3a, inset). When we investigated individual SF transcripts, we found that of the 167 well-characterized SF genes (Table 2, from [26]), 101 were alternatively spliced in B-ALL compared to pro-B cells (Fig 3b, bottom left, proB- vs B-ALL bar chart). Moreover, these changes were highly specific to the malignant phenotype and not to states of normal B-cell differentiation (Fig 3b, upper and middle panels). For example, only 6 SF transcripts were alternatively spliced during the early progenitors-to-proB transition and only 9 during the proB-to-preB transition (Fig 3b middle).

For each LSV, MAJIQ provides a ΔPSI (percent-spliced-in) value to indicate changes in isoform abundance. We observed that many members of the hnRNP and SR families, which play key roles in exon inclusion or skipping [27], exhibit profound changes in ΔPSI values, including increases in the so-called 'poison exons' with in-frame stop codons in several SRSF proteins [28] (Fig. 3c, left). On the other hand, hnRNPA1 showed increased skipping of exon 11, which was consistent among all leukemias analyzed (Fig. 3c, right). Of note, some transcripts contain multiple aberrant LSVs. For example, SRSF11 has three distinct LSVs with different corresponding start/end genomic coordinates.

## Dysregulated splicing of hnRNPA1 in B-ALL

According to the Ensembl database [29], there is evidence for hnRNPA1 transcripts with alternatively spliced 3' UTRs, notably the canonical long proximal exon 11 and two shorter, distal exons (exons 12/13, ENST00000547566) (Fig. 4a, top). We observed that the exon 11-containing transcript predominated in normal pro-B cells (Fig 4a, red), but in a typical B-ALL sample the predominant event was the skipping of exon 11 to exons 12 and 13 (Fig. 4a, blue). While exon 11 PSI values varied across leukemia samples, most leukemias had increased skipping of exon 11 compared to pro-B samples (Fig. 4b, blue). We also observed intron 10 inclusion in all pro-B and B-ALL samples, although this event was not significantly different between normal and malignant samples (Fig 4b, grey stacks). We validated and quantitated exon 11 inclusion by RT-qPCR using primers spanning exons 10 and 11. Using Spearman's *rho* statistic we observed a strong positive association between MAJIQ predictions and RT-qPCR validations (0.73, p-value<0.001) (Fig. 4c). We next applied the same correlation analysis to hnRNPA1 mRNA expression levels versus exon 11 inclusion and again found a positive correlation between the two measurements (0.43, p-value<0.01) (Fig. 4d). This suggested that preferential splicing from exon 10 to the distal 3' UTR exons 12/13 may decrease hnRNPA1 RNA steady state levels.

To model this event and identify potentially affected transcripts, we knocked down hnRNPA1 with siRNA in the Epstein-Barr Virus (EBV) transformed human B-lymphoblastoid P493-6 cell line [30] and performed RNA-seq analysis. Differential expression analysis showed that levels of hnRNPA1 mRNA (as well as its pseudogene P7) were robustly decreased by the hnRNPA1 siRNA compared to the non-targeting siRNA control (Fig. 4e). Then using MAJIQ we identified 213 LSVs in 184 genes associated with knockdown of hnRNPA1 (minimum of 20% ΔPSI, 95% CI). Of these hnRNPA1-dependent LSVs, 74 (or more than 30%) were present in at least one B-ALL sample (Fig. 4f), with more than half of these 74 LSVs altered in 10 or more samples. For example, one of the LSVs in the DICER1 gene was altered in B-ALL samples in the same manner as in the hnRNPA1 KD cells (Fig. 4g). Interestingly, this LSV maps to the 5'UTR of the DICER1 transcript, potentially diminishing its translation efficiency. This type of deregulation would be consistent with the loss-of-function mutation and copy number variations in DICER1 in several types of cancer including leukemias [31].

## Aberrant splicing of leukemia drivers in pediatric B-ALL

We next aimed to identify alternative splicing in other genes that contribute to leukemogenesis. We retrieved all genes (141) with known somatic mutations in hematologic malignancies from the COSMIC

database (v.82) [20, 21] (Table 3) and searched for aberrant LSVs affecting these genes in B-ALL samples using the very stringent ΔPSI>50% cutoff for exon inclusion/skipping. We identified 81 aberrant LSVs in 41 genes that were present in at least two B-ALL samples. They accurately separated leukemia samples from the non-transformed pro-B cell counterparts following hierarchical clustering using Euclidian distances (Fig. 5a). Of note, these LSVs affected roughly a third of B-ALL driver genes. Some (in genes such as FBXW7) were present in almost all B-ALL samples, attesting to their potential significance in leukemia pathogenesis.

To confirm our findings in independent datasets, we searched for aberrant LSVs in the COSMIC genes in the TARGET [32] and St Jude Children's Research Hospital (SJCRH) [33] B-ALL datasets using pediatric pro-B cells for comparison. We found 229 LSVs in 80 genes in the TARGET data and 362

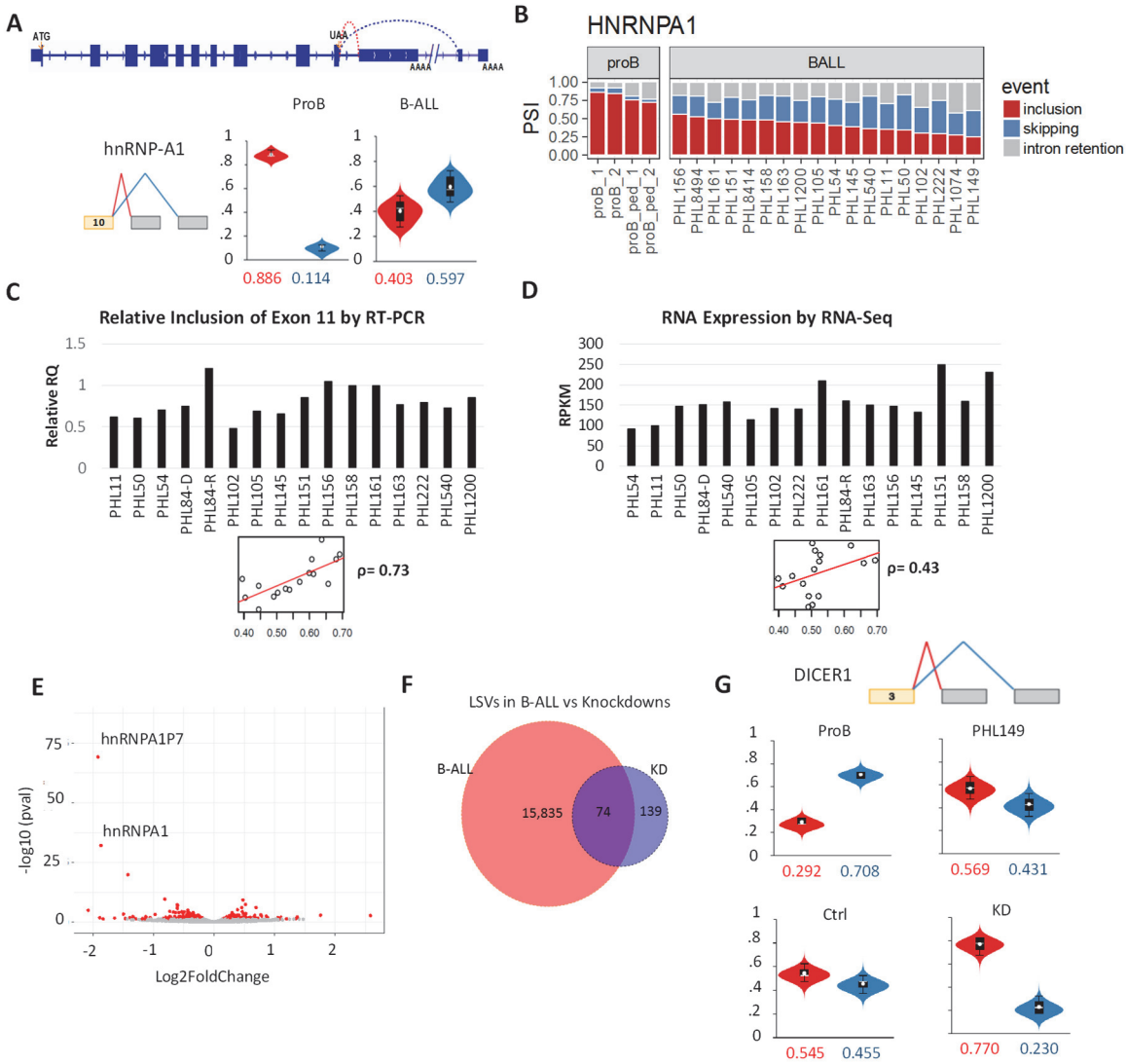


Fig. 4. Alternative splicing of hnRNPA1 in pediatric B-ALL. **a.** Exon-intron structure (top) and violin plots (bottom) depicting alternative splicing of the hnRNPA1 transcript. The violin plot on the right represents a representative B-ALL sample. **b.** Stack plot of PSI values of exon 11 corresponding to pro-B cell fractions and 18 B-ALL samples. **c.** Analysis of exon 11 inclusion in B-ALL samples by qRT-PCR. Spearman's correlations of RT-PCR vs. MAJIQ values are shown below. **d.** hnRNPA1 mRNA expression according to RNA-Seq. Spearman's correlations of RNA-Seq vs. RT-PCR are shown below. **e.** Volcano plot showing differential mRNA expression analysis in hnRNPA1 knockdown versus control cells. **f.** Overlap in LSVs between B-ALL and hnRNPA1 knockdown. **g.** Violin plots depicting alternative splicing of DICER1 in Pro-B vs. PHL149 B-ALL (top) and hnRNPA1 control and knockdown P493-6 cells (bottom).

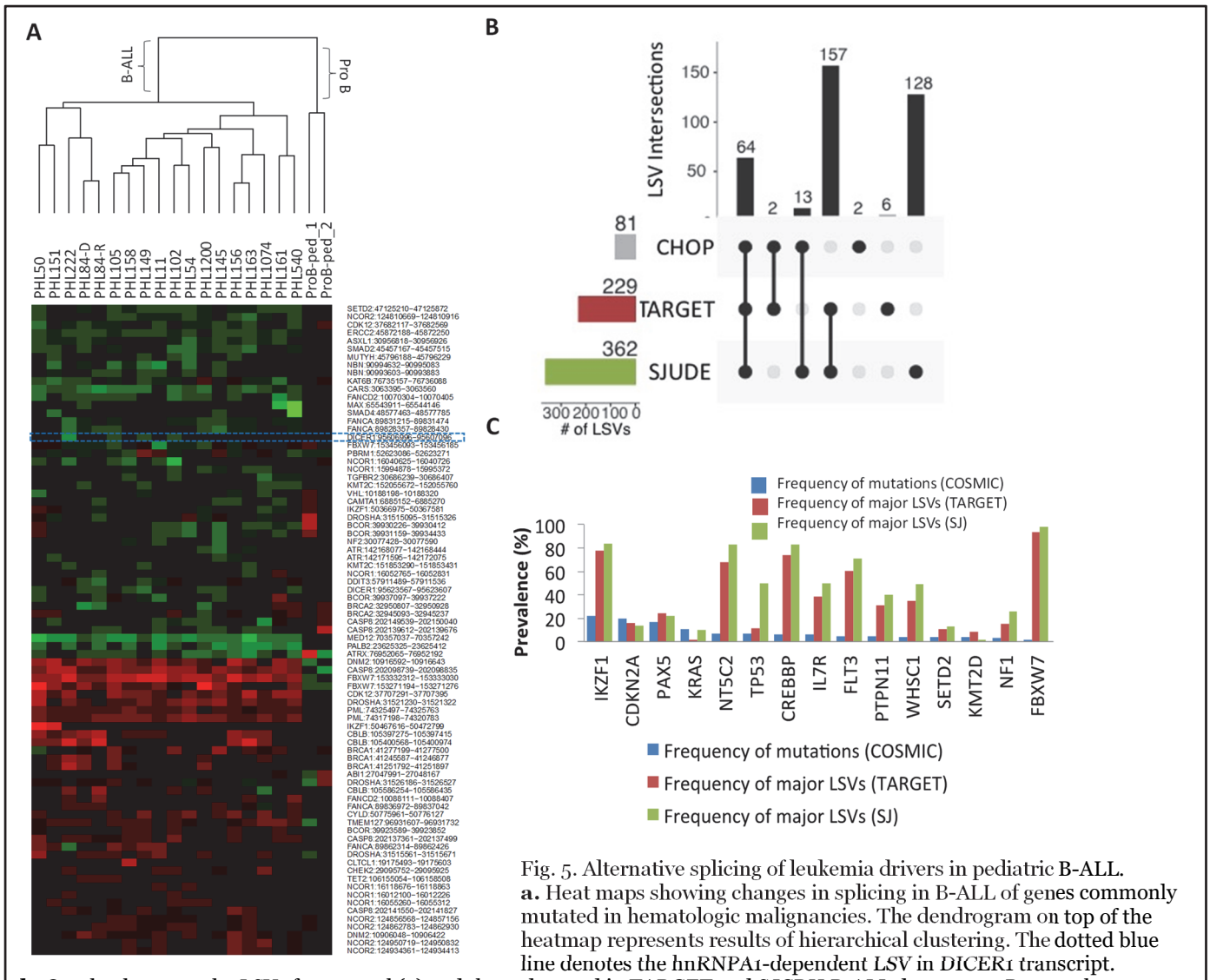


Fig. 5. Alternative splicing of leukemia drivers in pediatric B-ALL. **a.** Heat maps showing changes in splicing in B-ALL of genes commonly mutated in hematologic malignancies. The dendrogram on top of the heatmap represents results of hierarchical clustering. The dotted blue line denotes the *hnRNPA1*-dependent *LSV* in *DICER1* transcript. **b.** Overlap between the LSVs from panel (a) and those detected in TARGET and SJCRH B-ALL datasets. **c.** Bar graph representing frequencies of mutations (COSMIC dataset) and major LSVs as defined by MAJIQ in the indicated datasets.

LSVs in 95 genes in the SJCRH data (Fig 5b). Importantly, 64 of 81 LSVs identified in CHOP datasets significantly overlapped with TARGET (Fisher's Exact Test, p=7e-124) and SJCRH datasets (Fisher's Exact Test, p=1e-139), 13 different LSVs are present in both CHOP and SJCRH datasets and another 2 LSVs are present in CHOP and TARGET datasets (Fig 5b). We then narrowed our analysis to LSVs in top 20 B-ALL tumor suppressors and oncogenes as defined in the COSMIC database. In both datasets, we discovered largely overlapping LSVs corresponding to 15 out of 20 genes, such as IL7R, FLT3, TP53, etc. Of note, LSV frequencies were much greater than frequencies of somatic mutation (Fig. 5c), suggesting that in B-ALL these driver genes are preferentially affected by post-transcriptional mechanisms, such as alternative splicing.

## DISCUSSION

Personalized cancer diagnostics traditionally employ selected oncogene panels, which can identify mutations in specific genes known or suspected to be drivers in human malignancies. Hematologic malignancy sequencing panel typically include dominant oncogenes (e.g., FLT3 and IL7R), recessive tumor suppressors (e.g., TP53 and FBXW7), and haploinsufficient DNA/RNA caretakers (e.g., splicing factor SRSF2). Results from such genetic profiling of diagnostic cancer specimens can identify prognostic mutations and inform treatment selection for patients. Our data demonstrate that genetic deregulation occurs in B-ALL at the level of splicing in the absence of genetic mutations. For example, our analyses of B-ALL transcriptomes demonstrated that several SRSF genes controlling exon inclusion [27] show widespread variations in their own splicing patterns, some of which are known to decrease protein levels [28]. Consequently, we observed aberrant splicing of some of their known target transcripts such as TP53 [34], which encodes the key tumor suppressor gene. Thus, our current data show that clinical genetic testing panels may be inadequate to identify all potential therapeutic vulnerabilities within B-ALL cells. Of note, the majority of aberrant LSVs were highly concordant across different datasets. This reproducibility validates our conclusions and alleviates the potential concern that sample preparation conditions (e.g., storage at ambient temperature) could be affecting RNA surveillance and thus impacting analysis of alternative splicing [35].

One of the most consistent changes in exon usage we observed was non-canonical selection of 3' UTRs of hnRNPA1, a splice factor implicated in cancer progression and drug resistance. This event correlated with a decrease in hnRNPA1 mRNA abundance. While elucidation of the underlying mechanism was beyond the scope of this study, we speculate that new 3'UTR sequences could create additional sites for targeting by microRNAs, which are known to play a key role in hnRNPA1 downregulation in chemotherapy-resistant ovarian cancer cells [36]. Interestingly, knockdown of hnRNPA1 in an Epstein-Barr Virus (EBV) transformed human B-cell lymphoblastoid cell line resulted in aberrant splicing of Dicer, a key enzyme in microRNA biogenesis (reviewed in [37]). Beyond the hnRNPA1-Dicer axis, we identified strong (ΔPSI>50%) LSVs in ~30% of genes included in the COSMIC database because of their documented involvement in hematologic malignancies [20, 21]. Interestingly, four of these LSVs affect Drosha, another key enzyme in the microRNA biogenesis pathway [37], suggesting a strong connection between splicing and miRNA machineries.

Of even greater importance is the fact that B-ALL-specific LSVs affect 15 out of 20 top leukemia driver genes (including the aforementioned *FLT3*, *IL7R*, and *TP53*), with frequencies far exceeding those of somatic mutations. This discovery could explain why the prevalence of somatic mutations and copy number variations in B-ALL is low compared to other human cancers. It remains to be determined whether splicing alterations in oncogenes and tumor suppressor genes are functionally equivalent to gain-of-function and loss-of-function mutations, respectively. If so, interfering with splicing using RNA-based therapeutics and/or available small molecule inhibitors could be used to inhibit oncogenes such as *FLT3* and to activate dormant tumor suppressor gene such as *TP53*. Such strategies could yield tangible therapeutic benefits across a broad spectrum of childhood B-ALL subtypes.

***Table 1. Clinical and genetic characteristics of childhood B-ALL samples.***
*USI = unique specimen identifier, D = diagnosis, R = relapse, F= female, M = male, SR = standard risk, HR = high risk.*

| USI | Disease status* | Age at Dx, yrs | NCI Risk Status# | Sex | Race | Ethnicity | Major cytogenetic and molecular lesions |
|---|---|---|---|---|---|---|---|
| 11 | R | 7.5 | SR | F | Black or African American | Not Hispanic or Latino | high hyperdiploidy with +4 +10 |
| 50 | D | 11.7 | HR | M | White | Not Hispanic or Latino | 46,XY, IKZF1 deletion |
| 54 | D | 3.9 | SR | M | Other | Hispanic or Latino | ETV6-RUNX1, PAX5 deletion |
| 102 | D | 6.6 | SR | F | American Indian or Alaska Native | Not Hispanic or Latino | ETV6-RUNX1 |
| 84 | D | 16.1 | HR | F | Black or African American | Not Hispanic or Latino | high hyperdiploidy |
| 84 | R | | | | | | high hyperdiploidy, CREBBP deletion |
| 105 | D | 2.4 | SR | F | White | Not Hispanic or Latino | high hyperdiploidy with +4 +10 |
| 145 | D | 7.7 | HR | M | Other | Hispanic or Latino | EBF1-PDGFRB, IKZF1 deletion |
| 149 | D | 15.0 | HR | F | Other | Not Hispanic or Latino | 45,XX , CDKN2A/B deletion |
| 151 | D | 8.5 | SR | F | White | Not Hispanic or Latino | CDKN2A deletion, IKZF1 deletion, ERG deletion |
| 156 | D | 6.3 | HR | F | Other | Not Hispanic or Latino | high hyperdiploidy, iAMP21 |
| 158 | D | 3.5 | SR | F | White | Not Hispanic or Latino | ETV6-RUNX1 |
| 161 | D | 3.9 | HR | M | White | Not Hispanic or Latino | 47,XY +22, iAMP21, IKZF1 deletion |
| 163 | D | 5.6 | SR | M | White | Not Hispanic or Latino | ETV6-RUNX1, PAX5 deletion |
| 222 | D | 16.0 | HR | M | Black or African American | Not Hispanic or Latino | 46,XY, IKZF1 deletion, |
| 540 | D | 1.9 | HR | F | White | Not Hispanic or Latino | ETV6-RUNX1 |
| 1074 | D | 5.2 | SR | F | White | Not Hispanic or Latino | ETV6-RUNX1, CDKN2A/B deletion |
| 1200 | D | 5.6 | SR | M | White | Not Hispanic or Latino | ETV6-RUNX1, NRAS Q61H; germline +21 |

### *Table 2. Well characterized splicing factors analyzed for aberrant LSVs.*

| Gene names | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A1CF | ESRP2 | HNRNPH1 | MBNL1 | RALY | RBM42 | SART3 | SRSF3 | ZC3H10 |
| ANKHD1 | EWSR1 | HNRNPH2 | MBNL2 | RBFOX1 | RBM45 | SF1 | SRSF4 | ZC3H14 |
| CELF1 | FMR1 | HNRNPH3 | MBNL3 | RBFOX2 | RBM46 | SF3A1 | SRSF5 | ZCRB1 |
| CELF2 | FUS | HNRNPK | MSI1 | RBFOX3 | RBM47 | SF3B1 | SRSF6 | ZNF638 |
| CELF3 | FXR1 | HNRNPL | NOVA1 | RBM10 | RBM4B | SFPQ | SRSF7 | ZRANB2 |
| CELF4 | FXR2 | HNRNPLL | NOVA2 | RBM11 | RBM5 | SNRNP70 | SRSF8 | ZRSR1 |
| CELF5 | G3BP1 | HNRNPM | PABPC1 | RBM15 | RBM6 | SNRPA | SRSF9 | ZRSR2 |
| CELF6 | G3BP2 | HNRNPR | PABPC3 | RBM15B | RBM7 | SRP54 | SYNCRIP | |
| CNOT4 | GEMIN5 | HNRNPU | PABPC4 | RBM17 | RBM8A | SRPK1 | TARDBP | |
| CPEB1 | HNRNPA0 | HNRNPUL1 | PABPC5 | RBM20 | RBMS1 | SRPK2 | TIA1 | |
| CPEB2 | HNRNPA1 | HNRNPUL2 | PABPN1 | RBM22 | RBMS3 | SRPK3 | TIAL1 | |
| CPEB3 | HNRNPA1L2 | IGF2BP1 | PCBP1 | RBM23 | RBMX | SRRM1 | TRA2A | |
| CPEB4 | HNRNPA2B1 | IGF2BP2 | PCBP2 | RBM24 | RBMXL1 | SRRM2 | TRA2B | |
| DAZAP1 | HNRNPA3 | IGF2BP3 | PCBP3 | RBM25 | RBMY1A1 | SRRM3 | TUT1 | |
| ELAVL1 | HNRNPAB | KHDRBS1 | PPRC1 | RBM28 | RBMY1B | SRRM4 | U2AF1 | |
| ELAVL2 | HNRNPC | KHDRBS2 | PRPF40A | RBM3 | RBMY1D | SRSF1 | U2AF1L4 | |
| ELAVL3 | HNRNPCL1 | KHDRBS3 | PRPF40B | RBM38 | RBMY1E | SRSF10 | U2AF2 | |
| ELAVL4 | HNRNPD | KHSRP | PTBP1 | RBM39 | RBMY1F | SRSF11 | WT1 | |
| ENOX1 | HNRNPDL | LIN28A | PTBP2 | RBM4 | RBMY1J | SRSF12 | YBX1 | |
| ESRP1 | HNRNPF | MATR3 | QKI | RBM41 | SAMD4A | SRSF2 | YBX2 | |

### *Table 3. COSMIC genes mutated in hematologic malignancies and analyzed for aberrant LSVs.*

| Gene names | | | | | | | |
|---|---|---|---|---|---|---|---|
| ABI1 | BCOR | CDKN2A | EIF3E | FBXW7 | MSH2 | RB1 | ZFHX3 |
| ACVR2A | BLM | CDKN2C | EP300 | FH | MSH6 | RBM10 | |
| ALDH2 | BRCA1 | CDX2 | ERCC2 | FHIT | MUTYH | SDHA | |
| AMER1 | BRCA2 | CEBPA | ERCC3 | FLCN | MYH9 | SETD2 | |
| APC | BRIP1 | CHEK2 | ERCC4 | FUS | NAB2 | SFRP4 | |
| ARHGAP26 | BTG1 | CIITA | ERCC5 | GRIN2A | NBN | SH2B3 | |
| ARHGEF12 | BUB1B | CLTC | ETNK1 | HNF1A | NCOA4 | SMAD2 | |
| ARID1A | CAMTA1 | CLTCL1 | EXT1 | IKZF1 | NCOR1 | SMAD3 | |
| ARID1B | CARS | CNBP | EXT2 | KAT6B | NCOR2 | SMAD4 | |
| ARID2 | CASP8 | CNOT3 | FAM46C | KDM5C | NDRG1 | SOCS1 | |
| ASXL1 | CBFB | CREB3L1 | FANCA | KEAP1 | NF1 | STK11 | |
| ATM | CBLB | CTCF | FANCC | KLF6 | NF2 | TET2 | |
| ATP2B3 | CCDC6 | CYLD | FANCD2 | KMT2C | NFKBIE | TGFBR2 | |
| ATR | CCNB1IP1 | DDIT3 | FANCE | LRIG3 | NRG1 | TMEM127 | |
| ATRX | CD274 | DDX10 | FANCF | LRP1B | PALB2 | TPM3 | |
| AXIN1 | CDC73 | DDX3X | FANCG | LZTR1 | PBRM1 | TSC1 | |
| AXIN2 | CDH1 | DICER1 | FAS | MAX | PML | TSC2 | |
| B2M | CDH11 | DNM2 | FAT1 | MED12 | PTEN | VHL | |
| BAP1 | CDK12 | DNMT3A | FAT4 | MEN1 | PTPN13 | XPA | |
| BCL10 | CDKN1B | DROSHA | FBXO11 | MLH1 | PTPRT | XPC | |

## METHODS

### Bone Marrow Fractionation.

Isolated mononuclear cells and whole bone marrow aspirates were obtained, respectively, from the University of Pennsylvania Stem Cell and Xenograft Core facility and CHOP Hematopathology Laboratory. For pediatric bone marrow samples, mononuclear cells were isolated by spinning over Ficoll gradient, as described earlier [38]. Residual red blood cells were lysed with Ammonium Chloride Lysis buffer with gentle rocking at room temperature for 10 min. Cells were pelleted by spinning at 250 x g for 10 min at 4° C and washed once with cold PBS/2%FBS. Cells were resuspended in 1mL PBS/2%FBS and incubated with 500uL FC Block on ice for 10 min. Cells were stained with 1mL CD34-PE, 500uL CD19-APC, and 500uL IgM-FITC for 30 min on ice. Cells were pelleted at 1300RPM for 6 min at 4° C and washed twice in cold PBS. Cells were resuspended in 3mL PBS/2%FBS containing 1uL/mL of 0.1mg/mL DAPI. Cells were sorted 4-ways using MoFlo ASTRIOS directly into RLT Lysis buffer (Qiagen) at a ratio of 1:3.

### Primary B-ALL Samples Acquisition

24 primary pediatric B-ALL samples were obtained from the CHOP Center for Childhood Cancer Research leukemia biorepository. Mononuclear cells from fresh bone marrow or peripheral blood specimens were purified via Ficoll gradient separation [38] and cryopreserved for downstream experimental use.

### RNA-seq of bone marrow fractions, primary samples and cell lines

RNA was isolated using Qiagen RNeasy Mini Kit. RNA integrity and concentration were found using Eukaryote Total RNA Nano assay on BioAnalyzer (CHOP NAPCore). RNA-seq was performed on 10ng-1ug of total RNA according to GeneWiz protocol of PolyA selection, Illuminia Hi-seq, 2x150bp pair-end, 350M raw reads per lane.

### RNA-Seq alignment, quantification, and differential expression

Fastq files of RNA-seq obtained from GeneWiz were mapped using STAR aligner [39]. STAR was run with the option "alignSJoverhangMin 8". We generated STAR genome reference based on the hg19 build. Alignments were then quantified for each mRNA transcript using HTSeq with the Ensemble-based GFF file and with the option "-m intersection-strict". Normalization of the raw reads was performed using the trimmed mean of M-values (TMM). Differential expression of wild-type and knock-down RNA-Seq datasets were assessed based on a model using the negative binomial distribution, a method employed by the R package DESeq2 [40]. Subsequent volcano plots were generated using the R package ggplot2. Those differential genes that had a p-value of $<0.05$ were deemed as significantly up or down-regulated. Principal component analysis (PCA) was done on normalized count values of the samples using a correlation matrix and a calculated score for each principal component.

### Splicing analysis

In order to detect LSVs we used the MAJIQ (version 1.03) tool [23]. We ran MAJIQ on the Ensemble-based GFF annotations, disallowing *de novo* calls [29]. We chose for further analysis LSVs that had at least a 20% change at a 95% confidence interval between two given conditions. Using in-house customized Perl and Ruby scripts, we filtered for events that corresponded to exon inclusion events only, forcing one event per LSV. Heatmaps were generated for each ΔPSI value of each LSV that passed the 20% change threshold between Pro-B and B-ALL and additionally filtered for a frequency of $n>=2$ in our B-ALL samples. These were generated using the R package gplots.

### siRNA knock-down of hnRNPA1

Biological duplicate experiments were performed on 2 million P493-6 B-lymphoblastoid cells [30] electroporated using Amaxa Program O-006 with either 133nM non-targeting siRNA (Dharmacon) or

300nM ON-TARGET Plus Human hnRNPA1 SMARTpool siRNA (Dharmacon). Cells were plated in 2mL warm tetracycline-free RPMI for 24 hr. RNA isolation and RNA-seq were performed as described above.

### Datasets

Cancer gene symbols and annotations were downloaded from the COSMIC database [20, 21]. Known splice factors were annotated and obtained from published studies or ensemble-annotated databases.

Pediatric B-ALL samples from the TARGET consortium (phs000218.v19.p7) were accessed via NCBI dbGaP (the Database of Genotypes and Phenotypes) Authorized Access system. Pediatric B-ALL samples from the St Jude Children's Research Hospital (EGAD00001002704 and EGAD00001002692) were accessed by permission from the Computational Biology Committee though The European Bioinformatics Institute (EMBL-EBI).

## ACKNOWLEDGEMENTS

### Funding

This work was supported by grants from the NIH (T32 HL007439 to KLB, T32 CA 115299 to EG, and K08 CA184418 to SKT), Stand Up To Cancer - St. Baldrick's Pediatric Dream Team Translational Research Grant (SU2C-AACR-DT1113 to ATT), William Lawrence and Blanche Hughes Foundation (2016 Research Grant ATT), St. Baldrick's Foundation (RG 527717 to ATT), Alex's Lemonade Stand Foundation (Innovation Award to ATT), and CURE Childhood Cancer (Basic Research Award to KLB).

### Availability of data and materials

The original RNA-Seq data sets will be deposited in the NCBI GEO database (accession number pending)

### Authors' contributions

KLB, ASN, and ATT conceived and designed the research. KLB, EG and SYY performed B-cell purification, fractionation, and RNA isolation. KLB performed RNA isolation from primary B-ALL specimens. AB and VP assisted with sample acquisition and flow cytometry. ASN, KEH, MRG, DT, and YB conducted the bioinformatics analyses. SKT and MC compiled and analyzed clinical data. KWL and ATT participated in data interpretation. KLB, ASN, and ATT wrote the manuscript. All authors read and approved the final submission.

### Competing interest

The authors declare that they have no competing interests.

### Ethics approval

B-ALL samples used in this submission were passed to investigators with a coded identifier and no protected health information (PHI). Basic demographic, treatment, relapse, and survival outcome data for delivered specimens were provided through an online honest broker system. Specimens were obtained via informed consent on institutional research protocols in accordance with the Declaration of Helsinki.

## REFERENCES


1. Roberts KG, Mullighan CG: Genomics in acute lymphoblastic leukaemia: insights and treatment implications. *Nat Rev Clin Oncol* 2015, **12:**344-357.
2. Scheuermann RH, Racila E: CD19 antigen in leukemia and lymphoma diagnosis and immunotherapy. *Leuk Lymphoma* 1995, **18:**385-397.
3. Sikaria S, Aldoss I, Akhtari M: Monoclonal antibodies and immune therapies for adult precursor B-acute lymphoblastic leukemia. *Immunol Lett* 2016, **172:**113-123.
4. Topp MS, Gokbuget N, Stein AS, Zugmaier G, O'Brien S, Bargou RC, Dombret H, Fielding AK, Heffner L, Larson RA, et al: Safety and activity of blinatumomab for adult patients with relapsed or refractory B-precursor acute lymphoblastic leukaemia: a multicentre, single-arm, phase 2 study. *Lancet Oncol* 2015, **16:**57-66.
5. Maude SL, Frey N, Shaw PA, Aplenc R, Barrett DM, Bunin NJ, Chew A, Gonzalez VE, Zheng Z, Lacey SF, et al: Chimeric antigen receptor T cells for sustained remissions in leukemia. *N Engl J Med* 2014, **371:**1507-1517.
6. Gardner R, Wu D, Cherian S, Fang M, Hanafi LA, Finney O, Smithers H, Jensen MC, Riddell SR, Maloney DG, Turtle CJ: Acquisition of a CD19-negative myeloid phenotype allows immune escape of MLL-rearranged B-ALL from CD19 CAR-T-cell therapy. *Blood* 2016, **127:**2406-2410.
7. Jacoby E, Nguyen SM, Fountaine TJ, Welp K, Gryder B, Qin H, Yang Y, Chien CD, Seif AE, Lei H, et al: CD19 CAR immune pressure induces B-precursor acute lymphoblastic leukaemia lineage switch exposing inherent leukaemic plasticity. *Nat Commun* 2016, **7:**12320.
8. Maino E, Bonifacio M, Scattolin AM, Bassan R: Immunotherapy approaches to treat adult acute lymphoblastic leukemia. *Expert Rev Hematol* 2016, **9:**563-577.
9. Haso W, Lee DW, Shah NN, Stetler-Stevenson M, Yuan CM, Pastan IH, Dimitrov DS, Morgan RA, FitzGerald DJ, Barrett DM, et al: Anti-CD22-chimeric antigen receptors targeting B-cell precursor acute lymphoblastic leukemia. *Blood* 2013, **121:**1165-1174.
10. Raetz EA, Cairo MS, Borowitz MJ, Blaney SM, Krailo MD, Leil TA, Reid JM, Goldenberg DM, Wegener WA, Carroll WL, et al: Chemoimmunotherapy reinduction with epratuzumab in children with acute lymphoblastic leukemia in marrow relapse: a Children's Oncology Group Pilot Study. *J Clin Oncol* 2008, **26:**3756-3762.
11. Fry TJ, Shah NN, Orentas RJ, Stetler-Stevenson M, Yuan CM, Ramakrishna S, Wolters P, Martin S, Delbrook C, Yates B, et al: CD22-targeted CAR T cells induce remission in B-ALL that is naive or resistant to CD19-targeted CAR immunotherapy. *Nature Med* 2017**:**10.1038/nm.4441.
12. Sotillo E, Barrett DM, Black KL, Bagashev A, Oldridge D, Wu G, Sussman R, Lanauze C, Ruella M, Gazzara MR, et al: Convergence of acquired mutations and alternative splicing of CD19 enables resistance to CART-19 immunotherapy. *Cancer Discov* 2015, **5:**1282-1295.
13. Alderton GK: Immunotherapy: Skipping out epitopes. *Nat Rev Cancer* 2015, **15:**699-699.
14. Behjati S: Hiding from the enemy. *Science Transl Med* 2015, **7:**313ec193-313ec193.
15. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, et al: Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 2011, **478:**64-69.
16. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C, et al: Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts. *N Engl J Med* 2011, **365:**1384-1395.
17. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, Krysiak K, Harris CC, Koboldt DC, Larson DE, et al: Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet* 2012, **44:**53-U77.
18. Quesada V, Conde L, Villamor N, Ordonez GR, Jares P, Bassaganyas L, Ramsay AJ, Bea S, Pinyol M, Martinez-Trillos A, et al: Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet* 2012, **44:**47-52.
19. Wang LL, Lawrence MS, Wan YZ, Stojanov P, Sougnez C, Stevenson K, Werner L, Sivachenko A, DeLuca DS, Zhang L, et al: SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. *N Engl J Med* 2011, **365:**2497-2506.
20. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, et al: COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015, **43:**D805-811.

21. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: A census of human cancer genes. *Nat Rev Cancer* 2004, **4:**177-183.
22. Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, Girtman K, Mathew S, Ma J, Pounds SB, et al: Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* 2007, **446:**758-764.
23. Vaquero-Garcia J, Barrera A, Gazzara MR, Gonzalez-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y: A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* 2016, **5:**e11752.
24. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003, **4:**3.
25. Blake JA, Dolan M, Drabkin H, Hill DP, Li N, Sitnikov D, Bridges S, Burgess S, Buza T, McCarthy F, et al: Gene Ontology annotations and resources. *Nucleic Acids Res* 2013, **41:**D530-535.
26. Sebestyen E, Singh B, Minana B, Pages A, Mateo F, Pujana MA, Valcarcel J, Eyras E: Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res* 2016, **26:**732-744.
27. Matera AG, Wang Z: A day in the life of the spliceosome. *Nat Rev Mol Cell Biol* 2014, **15:**108-121.
28. Jumaa H, Nielsen PJ: The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation. *Embo J* 1997, **16:**5077-5085.
29. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al: Ensembl 2018. *Nucleic Acids Res* 2017.
30. Pajic A, Staege MS, Dudziak D, Schuhmacher M, Spitkovsky D, Eissner G, Brielmeier M, Polack A, Bornkamm GW: Antagonistic effects of c-myc and Epstein-Barr virus latent genes on the phenotype of human B cells. *Int J Cancer* 2001, **93:**810-816.
31. Foulkes WD, Priest JR, Duchaine TF: DICER1: mutations, microRNAs and mechanisms. *Nat Rev Cancer* 2014, **14:**662-672.
32. Hunger SP, Loh ML, Whitlock JA, Winick NJ, Carroll WL, Devidas M, Raetz EA, Committee COGALL: Children's Oncology Group's 2013 blueprint for research: acute lymphoblastic leukemia. *Pediatr Blood Cancer* 2013, **60:**957-963.
33. Gu Z, Churchman M, Roberts K, Li Y, Liu Y, Harvey RC, McCastlain K, Reshmi SC, Payne-Turner D, Iacobucci I, et al: Genomic analyses identify recurrent MEF2D fusions in acute lymphoblastic leukaemia. *Nat Commun* 2016, **7:**13331.
34. Tang Y, Horikawa I, Ajiro M, Robles AI, Fujita K, Mondal AM, Stauffer JK, Zheng ZM, Harris CC: Downregulation of splicing factor SRSF3 induces p53beta, an alternatively spliced isoform of p53 that promotes cellular senescence. *Oncogene* 2013, **32:**2792-2798.
35. Dvinge H, Ries RE, Ilagan JO, Stirewalt DL, Meshinchi S, Bradley RK: Sample processing obscures cancer-specific alterations in leukemic transcriptomes. *Proc Natl Acad Sci USA* 2014, **111:**16802-16807.
36. Rodriguez-Aguayo C, Monroig PDC, Redis RS, Bayraktar E, Almeida MI, Ivan C, Fuentes-Mattei E, Rashed MH, Chavez-Reyes A, Ozpolat B, et al: Regulation of hnRNPA1 by microRNAs controls the miR-18a-K-RAS axis in chemotherapy-resistant ovarian cancer. *Cell Discov* 2017, **3:**17029.
37. Sotillo E, Thomas-Tikhonenko A: Shielding the messenger (RNA): microRNA-based anticancer therapies. *Pharmacol Ther* 2011, **131:**18-32.
38. Tasian SK, Doral MY, Borowitz MJ, Wood BL, Chen IM, Harvey RC, Gastier-Foster JM, Willman CL, Hunger SP, Mullighan CG, Loh ML: Aberrant STAT5 and PI3K/mTOR pathway signaling occurs in human CRLF2-rearranged B-precursor acute lymphoblastic leukemia. *Blood* 2012, **120:**833-842.
39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013, **29:**15-21.
40. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014, **15:**550.