Long-read transcriptome sequencing of CLL and MDS patients uncovers molecular effects of *SF3B1* mutations

Alicja Pacholewska, Matthias Lienhard, Mirko Brueggemann, et al.

Genome Res. published online September 13, 2024
Access the most recent version at doi:10.1101/gr.279327.124

Long-read transcriptome sequencing of CLL and MDS patients uncovers molecular effects of *SF3B1* mutations

Alicja Pacholewska[1,2,*,+], Matthias Lienhard[3,*], Mirko Brüggemann[4,*], Heike Hänel[5], Lorina Bilalli[1], Anja Königs[1,2], Felix Heß[1,2], Kerstin Becker[6,7], Karl Köhrer[6], Jesko Kaiser[8], Holger Gohlke[8,9], Norbert Gattermann[10], Michael Hallek[2,11], Carmen D. Herling[11,12], Julian König[5], Christina Grimm[1,2,$], Ralf Herwig[3,$,+], Kathi Zarnack[4,$,+], Michal R. Schweiger[1,2,$,+]

[1] Institute for Translational Epigenetics, Faculty of Medicine, University of Cologne, 50931 Cologne, Germany.

[2] Center for Molecular Medicine Cologne (CMMC), Faculty of Medicine, University of Cologne, 50931 Cologne, Germany.

[3] Department of Computational Biology, Max Planck Institute (MPI) for Molecular Genetics, 14195 Berlin, Germany.

[4] Buchmann Institute for Molecular Life Sciences and Institute of Molecular Biosciences, Goethe University Frankfurt, 60438 Frankfurt, Germany.

[5] Institute of Molecular Biology, 55128 Mainz, Germany.

[6] Genomics & Transcriptomics Laboratory, Biological and Medical Research Center, Heinrich-Heine-University, and West German Genome Center, 40225 Düsseldorf, Germany.

[7] Cologne Center for Genomics (CCG), University of Cologne, 50931 Cologne, Germany.

[8] Institute for Pharmaceutical and Medicinal Chemistry, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

[9] Institute of Bio- and Geosciences (IBG-4: Bioinformatics), Forschungszentrum Jülich, 52428 Jülich, Germany

[10] Department of Haematology, Oncology and Clinical Immunology, University Hospital Düsseldorf, 40225 Düsseldorf, Germany.

[11] Department I of Internal Medicine, Center for Integrated Oncology Aachen-Bonn-Cologne-Düsseldorf, University Hospital Cologne, 50937 Cologne, Germany.

[12] Medical Clinic and Policlinic 1, Hematology, Cellular Therapy, Hemostaseology, and Infectious Diseases, University of Leipzig Medical Center, 04103 Leipzig, Germany

[*,$] Equal contribution

[+] Corresponding authors

Running Title: Long-read transcriptomics in CLL and MDS

Abstract

Mutations in splicing factor 3B subunit 1 (*SF3B1*) frequently occur in patients with chronic lymphocytic leukemia (CLL) and myelodysplastic syndromes (MDS). These mutations have different effects on the disease prognosis with beneficial effect in MDS and worse prognosis in CLL patients. A full-length transcriptome approach can expand our knowledge on *SF3B1* mutation effects on RNA splicing and its contribution to patient survival and treatment options. We applied long-read transcriptome sequencing (LRTS) to 44 MDS and CLL patients, as well as two pairs of isogenic cell lines with and without *SF3B1* mutations, and found > 60% of novel isoforms. Splicing alterations were largely shared between cancer types and specifically affected the usage of introns and 3' splice sites. Our data highlighted a constrained window at canonical 3' splice sites in which dynamic splice site switches occurred in *SF3B1*-mutated patients. Using transcriptome-wide RNA binding maps and molecular dynamics simulations, we showed multimodal SF3B1 binding at 3' splice sites and predicted reduced RNA binding at the second binding pocket of SF3B1^{K700E}. Our work presents the hitherto most complete LRTS study of the *SF3B1* mutation in CLL and MDS and provides a resource to study aberrant splicing in cancer. Moreover, we showed that different disease prognosis results most likely from the different cell types expanded during carcinogenesis rather than different mechanisms of action of the mutated SF3B1. These results have important implications for understanding the role of *SF3B1* mutations in hematological malignancies and other related diseases.

Keywords:

Long-read transcriptome sequencing; LRTS, chronic lymphocytic leukemia, CLL, myelodysplastic syndrome, MDS, Iso-Seq, differential splicing, iCLIP

Introduction

Splicing is a fundamental step in eukaryotic gene expression in which non-coding introns are removed from pre-messenger RNA (pre-mRNA) transcripts and exons are joined to form mature mRNAs. This intricate process is often disrupted in cancer, either by mutations in spliceosomal genes or by other mechanisms that affect normal splicing function (Yang et al. 2021; Seiler et al. 2018; Shiozawa et al. 2018; Quesada et al. 2012; Bradley and Anczuków 2023). In turn, aberrant splicing can lead to changes in the composition of expressed isoforms and the formation of new isoforms that alter the encoded proteins and can have far-reaching consequences for cellular function. One striking example of splicing alterations in cancer are mutations in the gene encoding the splicing factor 3B subunit 1 (*SF3B1*) that have divergent ramifications for treatment efficiency and prognosis (Rossi et al. 2011; Papaemmanuil et al. 2011). Somatic *SF3B1* mutations are frequently found in myelodysplastic syndrome (MDS, 20%), chronic lymphocytic leukemia (CLL, 15%), acute myeloid leukemia (3%), uveal melanoma (20%), cutaneous melanoma (4%), prostate cancer (1%) and in 2% of all breast, pancreatic and lung cancers (Bland et al. 2023).

In CLL patients, *SF3B1* mutations are typically subclonal and have been linked to disease progression and shorter survival (Wan and Wu 2013; Landau et al. 2015). On the other hand, *SF3B1* mutations in MDS patients have been associated with specific disease phenotypes that show erythroid dysplasia with ring sideroblasts and ineffective erythropoiesis (Malcovati et al. 2015). Unlike in CLL patients, a positive effect of the *SF3B1* mutation on survival has been observed in almost all groups of MDS patients, except those with excess blasts, for whom no significant effect has been observed (Malcovati et al. 2020). However, so far there is no explanation for the divergent ramifications of *SF3B1* mutations in CLL and MDS pathology.

Pre-mRNA splicing is a highly dynamic process, and the spliceosome undergoes several structural changes from E to A, B, and C complex during splicing. SF3B1 is part of the spliceosome and plays a critical role in 3’ splice site usage. As a subunit of the U2 small nuclear ribonucleoprotein complex (snRNP), SF3B1 is UV-crosslinked with the pre-mRNA on both sides of the branch point (BP)

adenosine in the A-complex, at nucleotide positions -6 and +5 (Gozani et al. 1996, 1998). The most common mutations in *SF3B1* accumulate in the HEAT (Huntington, Elongation Factor 3, PR65/A, TOR) domain at its C-terminus (Supplemental Fig. S1). The HEAT domain consists of 20 non-identical HEAT repeats that form the RNA binding interface (Cretu et al. 2016). These mutations are predicted to impact the N-terminal domain involved in complex formation with other splicing factors (Canbezdi et al. 2021). Mutations in SURP and G-patch domain containing 1 (*SUGP1*) mimic the splice alterations of mutant SF3B1 (Liu et al. 2020; Alsafadi et al. 2021) and mutations in *DHX15* partially recapitulate the splicing alterations of mutant SF3B1 (Zhang et al. 2022). Both proteins have been shown to bind less to mutated SF3B1 (Zhang et al. 2019).

Previous studies based on short-read RNA sequencing (RNA-seq) have reported alternative 3' splice site usage (3'AS) and intron retention (IR) as the most prominent splicing alterations in CLL and MDS patients with mutated *SF3B1* (Shiozawa et al. 2018; Wang et al. 2016; Tang et al. 2020; DeBoever et al. 2015; Kesarwani et al. 2017). The alternative 3' splice sites (referred to as AG') that were preferably used upon *SF3B1* mutation, are enriched at approximately 20 nucleotides (nt) upstream of the canonical splice sites (AG) (Obeng et al. 2016; Wang et al. 2016; Tang et al. 2020; DeBoever et al. 2015). This strong positional constraint suggested that the mutations impacted SF3B1 binding and BP recognition upstream of the 3' splice sites. Additionally, it was proposed that the mutation promotes the usage of otherwise inaccessible AG' within the RNA secondary structure (Kesarwani et al. 2017). Despite these hypotheses, the exact mechanism of the effect of mutations in *SF3B1* is still not resolved.

Here, we aimed to comprehensively characterize the effects of *SF3B1* mutations in cancer using long-read transcriptome sequencing (LRTS) and combined complementary data derived from MDS and CLL patients with isogenic cell lines.

Results

Long-read RNA sequencing expands patient transcriptome landscapes in divergent biological contexts

To investigate the effect of *SF3B1* mutations on splicing, we characterized the transcriptomes of three datasets: CLL patients, MDS patients with ring sideroblasts, and isogenic cell lines with or without somatic *SF3B1* mutations (Supplemental Fig. S2). In brief, we collected CLL cells or whole-blood samples from 19 CLL and 25 MDS patients, including eight CLL patients and 14 MDS patients with mutations in the SF3B1 HEAT repeat domain (Fig. 1A). These were complemented by two isogenic leukemia cell line pairs (K562 and Nalm6), both with *SF3B1*$^{wt/wt}$ and *SF3B1*$^{mut/wt}$. K562 cells originated from a patient with chronic myeloid leukemia (CML) and the *SF3B1*$^{mut/wt}$ cells carried K700E the mutation, whereas the Nalm6 cells originated from a patient with B cell acute lymphoblastic leukemia (B-ALL) and the *SF3B1*$^{mut/wt}$ cells carried the H662Q mutation. As controls, we further included B cells from six healthy donors (Supplemental Table S1). The RNA expression level of mutated *SF3B1* ranged from 14% to 52% (43% on average) in patients, 43% in K562, and 29% in the Nalm6 cell line (Supplemental Table S1). To detect complete transcript isoforms, we performed LRTS using Iso-Seq® (Pacific Bioscience). We reached a mean of 582,135 full-length non-chimeric reads per sample, cumulating in a total of 33,763,806 reads with an average length of 2,721 bp (Supplemental Table S1). Only 9% of the reads were potentially affected by technology-specific technical artefacts (Cocquet et al. 2006) (Supplemental Fig. S3–S4).

In total, we identified 89,659 substantially expressed transcripts that contributed to at least 1% to a gene's total expression and were covered by at least five full-length reads (Supplemental Fig. S5). Almost one third of these reads (28,261; 31.5%) were classified as full splice matches (FSM) to annotated isoforms. Moreover, 58,168 (64.9%) represented novel isoforms that only partially overlapped with gene annotations, and 3,230 (3.6%) reads originated from non-annotated, novel genes (Supplemental Fig. S6). Even for transcripts expressed at ≥ 1 transcript per million (TPM), the novel isoforms consisted of more than half of all transcripts detected (Fig. 1B). A large fraction of isoforms

was shared by the different patient cohorts and isogenic cell lines, with a larger overlap of expressed isoforms between *SF3B1*$^{mut/wt}$ and *SF3B1*$^{wt/wt}$ of the same dataset, than between datasets (Fig. 1C, Supplemental Fig. S7).

The *SF3B1* gene has multiple isoforms annotated and was indeed expressed in several isoforms in both, samples with or without *SF3B1* mutations (Fig. 1D, Supplemental Fig. S8). Although the most frequently expressed *SF3B1*-FSM isoform (around 70% of *SF3B1* transcripts) fully corresponded to the annotated isoform, two shorter novel isoforms showed a disease-specific expression almost exclusively in either CLL or MDS patients. These contributed approximately 10% each to the gene's overall expression, irrespective of the *SF3B1* mutational status (Fig. 1E). The CLL-specific isoform (*SF3B1*-CLL) showed retention of the fourth intron which introduced a premature termination codon and likely targeted the isoform for nonsense-mediated mRNA decay (NMD). In the MDS-specific isoform (*SF3B1*-MDS), the penultimate exon was skipped and induced a frameshift that probably resulted in an NMD-resistant isoform that encoded for a C-terminally truncated protein devoid of HEAT repeats 18–20 and the anchor domain. In addition to the divergent splicing pattern, *SF3B1* also showed three times higher expression in CLL compared to MDS patients, whereas its levels were reduced in MDS patients when compared to B cells from healthy donors (Supplemental Fig. S8).

Overall, our results demonstrated a high transcriptome information content in the patient cohorts, which was dominated by a large number of novel transcripts.

Patients and cell lines with *SF3B1* mutations show similar splicing defects

In order to investigate the transcriptome diversity at the splice-site level, we used the recently developed IsoTools software (Lienhard et al. 2023) to identify alternative splicing events (ASEs) in the transcripts expressed. IsoTools distinguishes exon skipping (ES), intron retention (IR), mutually exclusive exons (ME), and 5’ and 3’ alternative splice sites (5’AS and 3’AS) events, as well as alternative first and last junctions (AFJ and ALJ). Using a cut-off of $\geq$ 100 reads, ASEs were quantified as the proportion of reads supporting the ASE in relation to the sum of reads for all transcript isoforms, referred to as percent spliced-in index (PSI). This threshold is motivated from

extensive testing to optimize the detection of true splicing events while minimizing false positives and also ensures a sufficient number of individual samples (≥ 5) supporting the vast majority of ASEs (Supplementary Fig. S9). Across all samples, we discovered 80,995 ASEs in 9,746 genes, for which the less expressed ASE made up for at least 10% of the reads. For 75% of these events, at least one of the alternatives was not annotated (novel event) (Fig. 2A).

Next, we used IsoTools (Lienhard et al. 2023) to detect significant differences in splicing associated with *SF3B1* mutations. Because *SF3B1* mutations have been reported to convey either beneficial (MDS) or disadvantageous (CLL) effects on patient survival (Rossi et al. 2011; Papaemmanuil et al. 2011), we first tested for differential splicing in *SF3B1*$^{mut/wt}$ vs. *SF3B1*$^{wt/wt}$ samples, separately in each dataset. We detected 82, 288, and 219 ASEs in the isogenic cell lines, CLL, and MDS patients, respectively (adjusted p-value [q-value] with false discovery rate, FDR (Benjamini and Hochberg 1995) < 10%, Supplemental Table S2). Although we observed only a moderate overlap of the identified events between the datasets (Fig. 2B, Supplemental Fig. S10), to our surprise, the correlation of the PSI changes for the ASEs identified was high (Fig. 2C), indicating a common mutational effect. We found that the genes altered by the disease-specific ASE were generally expressed significantly higher in the corresponding group of patients (Fig. 2D, Supplemental Fig. S11) and out of the union of 531 genes with an ASE, 149 were relatively higher in MDS and 155 were higher in the CLL samples by at least 2-fold (FDR < 1%) based on the Iso-Seq data. Overall, about two thirds of ASEs called in the MDS or CLL dataset separately were significantly differentially expressed (FDR < 5%) when comparing MDS to CLL (Supplemental Fig. S12, Supplemental Table S3).

Our findings suggested that while *SF3B1* mutations introduced shared splicing effects in both CLL and MDS patients, the divergence in the disease outcome could be attributed to the differential transcriptomic profiles. Specifically, the mutation seemed to exert its most potent effects on genes that were already dominantly expressed in each disease.

SF3B1 mutations affect alternative 3' splice site usage and intron retention

Since the individual analyses for patients and isogenic cell lines indicated a common effect of *SF3B1* mutations, we combined all *SF3B1*$^{mut/wt}$ and *SF3B1*$^{wt/wt}$ samples to increase the statistical power for an overall estimation of the *SF3B1* mutational impact. In total, we identified 775 differential splicing events in 530 different genes (Fig. 3A, Supplemental Table S4). As reported previously (Shiozawa et al. 2018; Wang et al. 2016; Tang et al. 2020; DeBoever et al. 2015; Kesarwani et al. 2017; Darman et al. 2015), the splicing changes upon *SF3B1* mutation were strongly enriched for 3'AS (326, 42%) and IR (213, 27%) events which together accounted for more than two thirds of the significant changes. The majority of IR events showed decreased IR (89%), whereas the majority of 3'AS events (78%) showed higher PSI values, corresponding to longer exons in *SF3B1*$^{mut/wt}$ (Fig. 3B–C, Supplemental Fig. S13, Supplemental Table S4). Consistent with the common mutational effect, the regulated 3'AS events showed a uniform response across cohorts and allowed to cluster *SF3B1*$^{mut/wt}$ and *SF3B1*$^{wt/wt}$ samples (Fig. 3B).

Upon closer inspection, we found that two CLL-*SF3B1*$^{mut/wt}$ patient samples with mutations in the HEAT domain clustered with the *SF3B1*$^{wt/wt}$ samples when a subset of ASEs was used (208 ASEs with q-value$_{.}$ < 0.01). One of these patients carried a rare *SF3B1* mutation, Q699E, that had so far been reported only once, in a single patient with bladder urothelial carcinoma included in the TCGA Pan-Cancer Atlas (according to cBioPortal accessed on 2023.10.16). Moreover, overexpression of the SF3B1-Q699H construct in HEK293FT cells did not lead to any aberrant splicing, suggesting that this mutation is weakly pathogenic (Darman et al. 2015). The second patient carried two mutations: D894G in the HEAT repeat 11 (allele frequency (AF) = 51 %), and I704N in the HEAT repeat 6 (AF = 18%, Supplemental Table S1). Although the mutations in these two patients might have a weaker effect on the global splicing, we decided to keep the samples in our analysis to allow for more biological variance in the statistical analysis and detect stronger signals. For all other mutations, the position within the SF3B1 HEAT domain had no discernible influence on the clustering, indicating that the different mutations impair SF3B1 similarly (Fig. 3B). In fact, using general splicing information (PSI values), irrespective of regulation, we found that 3'AS events, but no other type of

ASEs, clearly differentiated the samples based on the *SF3B1* mutational status in an unsupervised principal component analysis performed on each dataset (Fig. 3C, Supplemental Fig. S14–S15), underlining the predominant effect of *SF3B1* mutation on 3'AS events.

In addition to many new differential ASEs, previously published discoveries could be confirmed, including e.g., five from 35 3'AS events (in *SEPTIN2*, *ERGIC3*, *RHNO1*, *FDPS*, and *SNRPN*) identified in CLL based on Nanopore sequencing (Tang et al. 2020), a 3'AS in *SEPTIN6* reported in MDS-*SF3B1*$^{mut/wt}$ patients (Dolatshad et al. 2016) (Supplemental Fig. S16), as well as thirteen 3'AS events (*BCL2L1*, *COASY*, *DPH5*, *DYNLL1*, *EI24*, *ERGIC3*, *MED6*, *METTL5*, *SERBP1*, *SKIV2L*, *TMEM14C*, *ZBED5*, and *ZDHHC16* that were consistently found in CLL, MDS, and uveal melanoma patients (Inoue et al. 2019; Pellagatti et al. 2018). Moreover, we confirmed 13 from 32 (Zhou et al. 2020) and 8 from 11 genes (Liu et al. 2020) reported as aberrantly spliced in either MDS or CLL patients.

The LRTS data opened the possibility to assess the splicing alterations in the context of complete transcript isoforms. Generally, we found that the effect of the *SF3B1* mutation on splicing did not influence the choice of transcript start or end sites, nor the probability of other splicing events of the same gene. This means that the effect was local, and the resulting alternative transcript corresponded to the canonical transcript, except for the single alternative event. This is exemplified by the *SF3B1*$^{mut/wt}$-induced inclusion of the poison cassette exon (PCE) in the *BRD9* gene that introduces a premature termination codon (PTC) (Inoue et al. 2019) (Supplemental Fig. S17–S18). Of note, the full-length reads allowed us not only to confirm differential splicing of this PCE, but also to locate it to a specific isoform that has not been annotated yet (NIC class). This long-read-derived novel isoform otherwise resembled the canonical *BRD9* isoform, whereas the annotated PTC isoforms were presumed to also harbor an alternative first exon and additional splicing alterations. Such incomplete and incorrect isoform annotations are likely to cause problems in quantifying transcriptome changes, especially when using short-read sequencing data.

Notably, we found multiple splicing factors among the genes affected by *SF3B1* mutations, Overrepresentation analysis revealed 18 from the spliceosome pathway (KEGG: 03040,

q-value 2.8×10^{-3}) (Supplemental Table S4). This indicated a broad impact of *SF3B1* mutations on the general splicing machinery, which could potentially lead to secondary effects on splicing. Taking a closer look at the 208 highly significant 3'AS events (q-value < 0.01), we identified four clusters based on PSI values detected in *SF3B1*$^{mut/wt}$ and *SF3B1*$^{wt/wt}$ (Fig. 3B, Supplemental Fig. S19). The cluster II with low PSI values in *SF3B1*$^{mut/wt}$ and no expression in *SF3B1*$^{wt/wt}$ was enriched in spliceosome (q-value = 0.0106) and cell cycle genes (GO: 0007049 q-value = 2.8×10^{-4}, Supplemental Fig. S20).

To independently validate the detected splicing changes, we performed short-read RNA-seq on the isogenic cell line pairs as well as on 27 CLL patient samples, which included the same 19 CLL samples used for Iso-Seq, and collected publicly available RNA-seq data from 398 MDS patients (Supplemental Table S1, Supplemental Fig. S21). Although we detected more events using rMATS (Shen et al. 2014) (Supplemental Fig. S22 bottom, Supplemental Table S4) we observed a high and significant (p-values < 0.001) correlation of PSI values in the ASEs detected with IsoTools (Lienhard et al. 2023) and rMATS (Shen et al. 2014) on the same cell lines (Pearson correlation coefficient R = 0.840, p-value = 2.47×10^{-17}). The same held true for the samples from CLL and MDS patients (R = 0.794 with p-value = 2.95×10^{-36} and R = 0.800 with p-value = 3.60×10^{-24}, respectively, Supplemental Fig. S23).

As an orthogonal approach, we employed semi-quantitative reverse transcription PCR (RT-PCR) to test 15 differential 3'AS events in the isogenic cell line pairs (Supplemental Fig. S24, Supplemental Table S5). From these 15 tested, 12 (80%) clearly showed an increase in alternatively spliced isoform expression in the *SF3B1*$^{mut/wt}$ conditions. The strongest effects were observed for 3'AS events in *MAP3K7*, *SEPTIN6*, and *SETD4*, which showed an almost complete switch to the alternative splicing variant (Fig. 3D–E). We additionally performed splicing reporter assays using minigenes for six 3'AS events in HEK293T cells (Fig. 3F). Indeed, we observed differential 3'AS usage upon ectopic *SF3B1*^{K700E} expression in four out of six minigenes tested (*SETD4*, *PRPF38A*, *THOC1*, and *SEPTIN2*, Fig. 3G), supporting that the effect of the *SF3B1* mutation persist in an unrelated cell line. In the two

remaining cases (*TPP2* and *BRCA1*) the usage of the upstream AG' was already low in the validation assays using the K562 and Nalm6 cell line pairs (Supplemental Fig. S24).

Overall, these results supported a common effect of *SF3B1* mutations in different biological backgrounds and confirmed their predominant impact on alternative 3' splice site usage and intron retention.

Computational prediction of protein function and stability of individual splicing isoforms.

Further on, we used the LRTS information on full-length transcripts to predict the potential coding sequences of the transcripts (Supplemental Methods). Among the 28,261 known transcripts (FSM), we found that 73.4% had a matching reference coding sequence (CDS) (Fig. 4A). In contrast, the majority of the 58,168 novel transcripts (89.7%) did not match a CDS and either lacked an open reading frame (ORF) (20.1%), initiated from an unannotated start codon (13.2%), or began at an annotated initiation site but deviated from the reference CDS due to alternative splicing (56.4%). Additionally, we observed that 35.1% of the expressed novel transcripts were likely to be targeted by NMD, compared to only 8.2% among the known transcripts (Fig. 4A).

We next determined how *SF3B1* mutation-induced alternative splicing impacts the coding potential and the function of the proteins. To this end, we classified ASEs into categories based on their relative location and the impact on the coding sequence: 5'UTR, disrupted start codon, in-frame, frameshift, disrupted stop codon, and 3'UTR. Of the 326 events featuring 3'AS, the majority led to either frameshift modifications (121 events) or in-frame changes (94 events) in the CDS (Fig. 4B). In total, we identified 274 ASEs predicted to yield stable alternative proteins which are not predicted to undergo NMD.

We further examined more thoroughly the functional consequences of the two novel *SF3B1* isoforms, one predominantly expressed in CLL and one in MDS (Fig. 1D–E). In the *SF3B1*-CLL transcript, the fourth intron was retained which contained a PTC that shortened the CDS to 465 nt. Our prediction showed a strong signal for NMD due to the presence of 19 downstream exon-exon junctions. In

contrast, in the *SF3B1*-MDS transcript, the penultimate exon was skipped, a frameshift was introduced and, subsequently, a PTC. However, because this PTC was located within the last exon, the *SF3B1*-MDS transcript was unlikely to be targeted by NMD and should result in a protein product missing its C-terminal section i.e., HEAT domains 18–20 and the terminal anchor domain (Fig. 4C).

To further investigate the functional outcomes of the altered proteins, we examined the impact on protein domain levels, by aligning Pfam (Mistry et al. 2021) domains to the predicted protein sequences. For 57 of the ASEs, we found at least one Pfam domain that overlapped the divergent part of the protein sequence, indicating partially altered protein functions (Fig. 4D, Supplemental Table S4) and we found evidence that some aberrantly spliced transcripts may induce a translations re-initiation event (Fig. 4E, Supplemental Fig. S25). We show this as an example for *MAP3K7*, a frequently described gene with an alternative splicing event in *SF3B1*$^{mut/wt}$ (Shiozawa et al. 2018; Wang et al. 2016; DeBoever et al. 2015).

The effect of *SF3B1* mutations depends on the distance and sequence context of alternative 3' AS splice sites

When we plotted the fraction of differential 3'AS against the splice site differences, we noticed, consistent with previous findings, that the alternative splice sites of differential 3'AS events were enriched within 12–21 nt upstream of the canonical splice site (AG) mainly used in the *SF3B1*$^{wt/wt}$ samples (Fig. 5A) (Obeng et al. 2016; Wang et al. 2016; Tang et al. 2020; DeBoever et al. 2015; Kesarwani et al. 2017). Within this range, 30.8% of 3'AS were significantly differentially used in *SF3B1*$^{mut/wt}$ compared to 1.8% outside this range. To investigate the significance of the AG'–AG distance and sequence context for the differential splicing event we constructed a minigene assay with a part of the *THOC1* transcript harboring the significantly affected 3'AS. The insert was then modified by replacing the 21 nt fragment between the AG' and AG of *THOC1* with 45–50 nt long AG'–AG fragments from alternatively but non-differentially spliced introns (*PABCL1*, *USP1*, *ZNF124*, Supplemental Fig. S26). As a control, we mutated AG' to GG', to disrupt any alternative splicing. (Fig. 5B). These experiments suggested that increasing the AG'–AG distance was sufficient to remove the 3'AS from SF3B1 regulation. We also found that specific sequences, such as the branch

point region upstream of AG', were responsible for the differential splicing between *SF3B1*^{wt} and *SF3B1*^{K700E} expressing cells and confirmed that a strong AG is required for AG' usage (Darman et al. 2015) (Fig. 5C, for details please see Supplemental Notes).

Our results confirmed that mutations in *SF3B1* primarily affected proximal AG', but the AG'–AG distance did not seem to be the sole factor required for the usage of AG'. Moreover, we did not find any motif enriched at this position that could indicate a binding of another protein potentially disrupting or competing with SF3B1. We therefore speculated that SF3B1 binding at the sites with ASE may be altered in patients carrying *SF3B1* mutation.

K700E mutation may lead to destabilization of SF3B1-mRNA binding

To scrutinize the effect of the most common *SF3B1* mutation, K700E, we performed molecular dynamics (MD) simulations of 20 transcripts with a 3'AS within 50 nt distance, including 14 transcripts which were differentially spliced between *SF3B1*$^{mut/wt}$ and *SF3B1*$^{wt/wt}$ and 6 non-differentially spliced transcripts (Supplemental Methods). For each transcript, we performed four replicas of 200 nt long MD simulations of the mRNA with: i) the BP of the downstream AG (BP) bound to SF3B1^{wt}; ii) the BP of the upstream AG (BP') bound to SF3B1^{wt}; iii) the BP bound to the SF3B1^{K700E} mutant; and iv) the BP' bound to the SF3B1^{K700E} mutant. There were no differences in BP binding between all transcript–protein combinations in the first binding pocket (Supplemental Fig. S27A–C, whereas the frequency of the interaction in the second binding pocket decreased in SF3B1^{K700E} due to repulsive interactions of like-charged atoms (Supplemental Methods, Supplemental Fig. S27A, D–E).

To investigate whether this decrease in interactions is associated with an increase in mRNA mobility, we computed the per-residue root mean square fluctuations (RMSF) of the mRNA nucleobases. Indeed, the mobility of nucleobases significantly increased for nucleobases in the 3' direction after the K700 binding site for SF3B1^{K700E} compared to SF3B1^{wt}, when the BP' was bound to SF3B1 (Supplemental Fig. S27F and Supplemental Fig. S28–S32).

Taken together, these results indicated that the K700E mutation did not lead to differences in the BP recognition. However, the mutation led to significant differences in SF3B1-mRNA contacts within the second mRNA-binding pocket at least for the 20 mRNAs tested with an upstream alternative AG within 50 nt.

SF3B1 shows multimodal binding at 3' splice sites

Our splicing analyses showed that mutations in *SF3B1* resulted in the activation of alternative splice sites 12 to 21 nt upstream from the canonical AG (Fig. 5A, 6A). To understand how SF3B1 recognizes these sites, we performed individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) to map SF3B1 binding sites throughout the transcriptome (König et al. 2010). After UV crosslinking, we immunoprecipitated SF3B1 from K562-SF3B1$^{wt/wt}$ and K562-SF3B1$^{K700E/wt}$ cells, yielding together more than 100 million SF3B1 crosslink events (Fig. 6B**,** Supplemental Table S1). To facilitate direct comparisons, we randomly subsampled the sequencing reads to adjust the library size of the replicates (see Methods). Based on the merged iCLIP data we identified 96,852 SF3B1 reproducible binding sites with an optimal width of 5 nt (Fig. 6C–D). The binding sites occurred in 8,127 genes, with the vast majority being protein-coding genes (93%). As expected, within the protein coding transcripts, SF3B1 mostly bound introns (94%, Fig. 6E–F).

Since K562-SF3B1$^{K700E/wt}$ cells expressed both wild-type and mutated SF3B1 protein and both variants were recognized by the anti-SF3B1 antibody that specifically binds to the SF3B1 N-terminus, we tested for differences in the SF3B1 binding between K562-SF3B1$^{K700E/wt}$ and K562-SF3B1$^{wt/wt}$. Consistent with a recent study (Porter et al. 2021), the K700E mutation did not generally impair RNA binding (Supplemental Fig. S33). Moreover, at the level of binding sites, we detected only minor differences between K562-SF3B1$^{K700E/wt}$ and K562-SF3B1$^{wt/wt}$ (Fig. 6G). Thus, although subtle differences may have been masked by the overlay of both protein variants in the heterozygous cells, the K700E mutation does not obviously change the global RNA binding behavior of SF3B1. However, local changes as predicted with the molecular dynamics simulations might be too dynamic to be caught by the global iCLIP analysis.

Next, we examined SF3B1 binding at 3' splice sites. Using meta-profiles, we detected two prominent peaks of SF3B1 binding (Fig. 6H). The two peaks were centered at about -50 nt and -10 nt upstream of the canonical 3' splice site and surrounded the BP (Fig. 6H). Visual inspection indicated multiple SF3B1 binding sites within each peak (Supplemental Fig. S34). When centering the metaprofiles to the predicted branch point adenosine, SF3B1 binds about 25 nt upstream and directly downstream of the predicted branch point adenosine (Fig. 6H). The binding peak at -10 nt of the canonical 3'splice site, coincided with the Py-tract region bound by U2AF2 (Zarnack et al. 2013). To test this, we performed iCLIP experiments with U2AF2 in K562-SF3B1$^{wt/wt}$ cells which confirmed that the -10 nt SF3B1 peak overlapped with U2AF2 binding (Fig. 6H). Together, these observations indicated that SF3B1 binds at both sites of the BP and that the binding site directly downstream of the BP encompasses the Py-tract.

Consistent with the two peaks of SF3B1 binding at 3' splice sites, we found that SF3B1 binding sites frequently occurred at distances of ~30 nt to each other (Fig. 6I). To globally classify the SF3B1 binding pattern, we merged adjacent binding sites into equal-sized binding regions (80 nt, 56,224 regions, Supplemental Table S6) and performed unsupervised uniform manifold approximation and projection (UMAP) followed by density-based of applications with noise (DBSCAN). This yielded three distinct clusters: Cluster C1 harbored mostly isolated binding sites (35,907 regions), cluster C2 included two closely spaced binding sites (5,635 regions), whereas cluster C3 showed a more complex arrangement of 3–4 binding sites with wider spacing (13,847 regions) (Fig. 6J–K, Supplemental Fig. S35). The latter were located closest to 3' splice sites (Fig. 6L), suggesting that multiple SF3B1 binding sites assemble into complex binding patterns at 3' splice sites. We then investigated the differences in binding between K562-SF3B1$^{K700E/wt}$ and K562-SF3B1$^{wt/wt}$ solely in the C3 cluster. We noticed that K562-SF3B1$^{K700E/wt}$ showed a slight decrease in the left peak, that was more distal to the canonical AG (Supplemental Fig. S36). Within cluster C3, we noticed a slight decrease in binding of K562-SF3B1$^{K700E/wt}$ compared to K562-SF3B1$^{wt/wt}$ specifically in the left peak, that was more proximal to the canonical AG (Supplemental Fig. S36). This suggested that the K700E

mutation led to change in the complex arrangement of SF3B1 binding sites at 3' splice sites, preferentially affecting AG-proximal binding.

Altogether, our SF3B1 iCLIP data showed that SF3B1 adopted a multimodal mode of binding at 3' splice sites, with two major peaks of SF3B1 binding that surround the BP. The peaks include multiple binding sites, which may reflect the dynamic binding rearrangements during the splicing process. The strong enrichment of this binding pattern at 3' splice sites suggested that the defined arrangement of binding is required for SF3B1's function in splicing.

SF3B1 alternates within a small window of alternative splice site distances and the K700E mutation leads to the use of the proximal upstream AG

We and others (Alsafadi et al. 2016; DeBoever et al. 2015; Darman et al. 2015; Zhang et al. 2019) found that 3' splice sites are particularly sensitive to *SF3B1* mutations when they are directly preceded by an alternative 3' splice site. To test how this relates to binding, we overlaid the iCLIP data with the splicing quantifications from the same isogenic cell lines (K562-SF3B1$^{wt/wt}$ and K562-SF3B1$^{K700E/wt}$). As shown above, the *SF3B1* mutations showed most prominent effects on alternative 3' splice sites 12–21 nt upstream of the canonical 3' splice site (Fig. 5A). At this distance, the upstream AG was at the border of the proximal peak of SF3B1 binding (Fig. 7A). Moving further away, the effects subsided drastically as soon as the upstream AG emerged from the proximal peak (Fig. 7B).

The 3'AS events within the critical window showed a distinct behavior already in wild-type K562-SF3B1$^{wt/wt}$ cells. If the distance between 3' potential splice sites was either < 12 nt or > 21 nt, splicing predominantly occurred at the upstream AG in the vast majority of cases, indicating that the spliceosome generally favored upstream AG usage. However, within the critical window, this pattern was inverted, such that upstream AG at these distances were generally outdone by the canonical downstream AG. This indicated that in presence of two AGs within 12–21 nt the downstream AG is predominately used for splicing. In line with this notion, we found that upstream AGs were generally depleted from this region (Fig. 7C, top panel). Moreover, upstream AG still

419 falling into this window showed basal usage already in wild-type cells in more than 50% of cases and
420 were predominantly activated upon *SF3B1* mutation (Fig. 7C, middle panel), which could be
421 facilitated by a less stable binding of SF3B1^{K700E} to the pre-mRNA as suggested by the molecular
422 dynamics model.

423 Based on these observations, we hypothesized that diminished SF3B1$^{K700E/wt}$ binding to the pre-
424 mRNA results in increased upstream AG usage within a critical window that may particularly impair
425 splicing fidelity. Indeed, *SF3B1* mutation preferentially affected the first (nearest) upstream AG (Fig.
426 7D) which lay significantly closer to the 3' splice site compared to non-differential 3'AS (Fig. 7E). In
427 contrast, the BP of the differential spliced 3'AS events were neither moved closer to the 3' splice site
428 nor differed in their predicted strengths (Fig. 7E, Supplemental Fig. S37), indicating that the
429 positioning of the upstream AG rather than the BP was the primary determinant for the observed
430 effects. The differentially used upstream AG had slightly lower predicted splice site strengths
431 (MaxEnt score (Yeo and Burge 2003)) than non-differential 3'AS. The downstream AGs were
432 considerably stronger than unused upstream AGs (Fig. 7C, lower panel). This was accompanied by
433 higher splice site strengths of the canonical 3' splice sites, indicating that a strong canonical 3' splice
434 site was required to support differentially spliced 3'AS events (Fig. 7F).

435 Taken together, we propose that SF3B1 binding often directly overlaps with alternative AG' in a
436 constrained window upstream of 3' splice sites (> 20 % Fig. 5D), thereby using downstream 3' splice
437 sites in wild-type conditions. In the presence of SF3B1 mutations, SF3B1 cannot properly bind the
438 Py-tract of the downstream canonical AG when a strong upstream AG' lies within a short distance,
439 leading to increased alternative splicing with the use of an alternative branch point (Supplemental Fig.
440 S38). This is partly explained by changes in the second binding pocket of SF3B1, as predicted by
441 molecular dynamics simulations, that destabilize the SF3B1-mRNA interaction. Thus, although
442 mutated SF3B1 still binds to both sides of the BP in this scenario, the changed SF3B1 protein
443 structure reduces usage of the downstream AG, resulting in widespread splicing defects.

Discussion

Alternative splicing plays a critical role in generating transcriptome diversity, and its dysregulation has been linked to various diseases, including cancer. Yet, complex transcriptome studies are challenging to perform with classical RNA-seq due to the frequent ambiguity in mapping short reads and the difficulties in identifying novel isoforms (Hooper 2014; Rehrauer et al. 2013; Zhang et al. 2017). Of note, our capacity to study transcriptomes and alternative splicing events has expanded with the recent advancements in long-read sequencing technologies. To the best of our knowledge, up to now only few studies applied long-read Oxford Nanopore sequencing to analyze splice site alterations mediated by mutated SF3B1 in CLL (bulk (Tang et al. 2020) or single cell (Peng et al. 2024)) and MDS (single-cell (Cortés-López et al. 2023)).

Here, we used long-read Iso-Seq (PacBio) sequencing of 44 patients to investigate the impact of *SF3B1* mutations on alternative splicing. Our LRTS analysis revealed a wide variety of transcripts, with more than two thirds of yet unannotated, or even over 3,000 transcripts from novel genes. This highlighted the importance of long-read sequencing for comprehensive transcriptome profiling, particularly for detecting novel splice variants and alternative splicing events. Our results supported previous findings that *SF3B1* mutations specifically alter the usage of 3' alternative splice sites and intron retention (DeBoever et al. 2015; Darman et al. 2015; Alsafadi et al. 2016). Importantly, using a comprehensive setup of two cohorts of CLL and MDS patients, complemented by two isogenic cell line pairs we were able to substantially expand the catalog of differentially spliced 3'AS to a total of 326 3'AS events in 266 genes (Supplemental Table S4). We observed similar effects in both patient cohorts and the cell lines studied, indicating a common effect of *SF3B1* mutation on splicing. The clinical differences and prognosis of the CLL and MDS patients with *SF3B1* mutations most likely depend on the specific gene expression profiles and thus their different relevance of splicing alterations.

Our results revealed that *SF3B1* mutations affect genes involved in the major mRNA splicing pathway, indicating a broader impact on the splicing machinery. This may trigger secondary effects

on splicing, potentially leading to altered transcriptomes and disease phenotypes. In particular and in concordance with previous studies, we observed an enrichment of the 3'AS events 12–21 nt upstream of the canonical 3' splice sites (DeBoever et al. 2015; Darman et al. 2015; Alsafadi et al. 2016). This region is typically depleted of alternative AG dinucleotides. Why AGs are depleted in the critical region remains unclear. One possibility is that they are removed by purifying selection and might have an evolutionary advantage. However, a subset of introns still contains AGs within this critical region leading to alternative 3'AS. Why some AGs remain present in the critical region is also unclear. We and others (Darman et al. 2015) observed that in these cases, the canonical AG is often stronger than at other 3' splice sites, indicating that a strong 3' splice site may be required to tolerate an upstream AG within the critical window.

Our minigene assays did not confirm that these 3'AS events depend solely on the distance between canonical and alternative AGs. Coinciding with the critical range of AG'–AG distances of 12–21 nt, we observed a bimodal binding of SF3B1 surrounding the BP, whereby the BP often coincided or was near to the upstream AG'. Thereby, SF3B1 binding may shield the upstream AG from recognition during splicing as was proposed by (Kesarwani et al. 2017). We did not detect obvious global changes in $SF3B1^{K700E}$ binding, which might be due to a temporary effect and/or binding affinity of the $SF3B1^{K700E}$ that might not be caught by iCLIP analyses. However, our MD simulations predict that the number of contacts between the mRNA and residue 700 of SF3B1 resulted in increased mobility of the mRNA at both, canonical and alternative 3' splice sites in $SF3B1^{K700E}$.

We confirmed with minigene assays that $SF3B1^{mut}$ uses an alternative BP that leads to 3'AS usage in $SF3B1^{mut/wt}$ cells (Darman et al. 2015; Alsafadi et al. 2016). However, transcriptome-wide studies revealed that about one third of all human exons have multiple branch points (Pineda and Bradley 2018; Mercer et al. 2015), which argues against the hypothesis that mutated SF3B1 always prefers usage of an alternative BP. Consistently, we did not observe any strong alterations in the binding of $SF3B1^{K700E/wt}$ to mRNAs, although a slight increase of the peak at the Py-tract directly upstream of the 3' splice site was observed (Supplemental Fig. S39). Whereas we cannot exclude that we partially co-precipitated U2AF2, we and others did not find obvious changes in U2AF2 binding to SF3B1 in

immunoprecipitations (Alsafadi et al. 2016; Cretu et al. 2016). Furthermore, U2AF2 binds to the Py-tract only during early stages of the splicing process and is released during transition to the activated B complex (Agafonov et al. 2011). In contrast, SF3B1 is assembled into the spliceosome later, and subsequently replaces U2AF2 in the activated B complex and binds to the Py-tract with its binding pocket consisting of HEAT domains 3–7 that harbor most of the mutational hotspots (Schmitzová et al. 2023). Therefore, the downstream peak observed in our SF3B1 iCLIP experiments most likely corresponds to SF3B1 binding.

The differential splicing observed in *SF3B1*$^{mut/wt}$ may result from a weakened binding of SF3B1^{mut}-containing spliceosomes. There might be an additional effect of decreased binding of DDX46/PRP5, a kinase involved in proof reading of the pre-mRNA branch site (Carrocci et al. 2017; Tang et al. 2016; Zhao et al. 2022; Zhang et al. 2021). Other splicing proteins that have been shown to bind less to SF3B1^{mut}, are DDX42 (Zhao et al. 2022), DHX15 (Zhang et al. 2024) and SUGP1 (Zhang et al. 2019, 2023). DDX42 and DDX46 have been shown to sequentially occupy the RNA binding pocket consisting of HEAT repeats 3–7 during early steps of the splicing process (Zhang et al. 2021, 2024; Yang et al. 2023).

Besides gaining additional insight into the SF3B1 splicing mechanism, we also explored the splicing and expression alterations identified through the Iso-Seq sequencing approach in CLL and MDS. Apart from identification of large numbers of new differentially spliced genes, we were able to specifically map the toxic exon of *BRD9* to its isoform and predict its amino acid composition. We also identified *SF3B1* isoforms specifically more present in MDS or CLL, albeit their expression levels were at approximately 10%. The S*F3B1*-CLL transcript is predicted to undergo NMD, whereas the *SF3B1*-MDS is predicted to miss its C-terminal part. This shortened protein might rescue part of the detrimental effect of the mutated *SF3B1* leading to a slightly favorable prognosis. *SF3B1* overexpression in CLL B cells with respect to normal B cells has been reported before (Wan and Wu 2013). Possible reasons for this increase might be: i) regulatory mechanisms, such as alterations in transcription factors or epigenetic changes; ii) oncogenic pathways involving growth factors or cytokines; or iii) feedback mechanisms (Huang et al. 2011). We would speculate that due to the

NMD-sensitive *SF3B1*-CLL transcript the cell increases the transcription of the regular version of the *SF3B1* transcript to compensate for this loss. This is not necessary in the case of MDS, because the *SF3B1*-MDS transcript is functional and only misses its C-terminal part. This would explain how the *SF3B1*-CLL isoform under NMD is related to the overexpression of *SF3B1* in CLL.

This overexpression can now be brought in context with the worse prognosis of mutated SF3B1 in CLL, because in the presence of the mutation this overexpression leads to a more massive dysregulation of alternative splicing caused by SF3B1 in CLL (compared to MDS) which in turn disrupts multiple pathways and lowers survival of the patients. Indeed, if we further investigate the disease specific ASEs in mutated vs wild-type *SF3B1* patients (Fig. 2B; Supplemental Table S2) we observed that the overall number of ASEs is fairly similar (CLL: 288, MDS: 219) with 1.31-fold increase in ASEs in CLL, but there was a drastic (× 2.75) increase of CLL-specific IR events (CLL: 77, MDS: 28) (Supplemental Fig. S40A). The ASEs break down to 69 (CLL) and 27 (MDS) unique genes and were largely different with only six genes in common (Supplemental Fig. S40B). The pathways affected by CLL-IRs were related to mRNA splicing machinery, oncogenic signaling pathways, as well as immune pathways that might affect the survival of the patients (Supplemental Fig. S40C). Thus, we would argue that the overexpression of *SF3B1* in CLL compared to MDS leads to elevated splicing effects, in particular introns, that target a different, more signaling-related panel of genes with multiple cellular functions that promote tumorigenesis and reduce survival. However, further functional analyses will show the impact of these altered SF3B1 proteins and if they influence MDS and CLL pathomechanisms.

Another example, which is frequently reported to be differentially spliced within *SF3B1* mutated cancers is *MAP3K7*. We were able to show that mutations in *SF3B1* led to reduced expression of longer isoforms and increased expression of isoforms with shortened protein kinase domain, likely impacting its function. Thus, with this data at hand it is possible to not only identify splicing events but to also map them to their cognate isoform and thus to provide information on the resulting protein composition. This is a fundamental information for understanding splicing data and to gain insight into pathomechanisms underlying CLL and MDS.

Our study provides new insights into the mechanism by which *SF3B1* mutations affect splicing regulation, and the potential consequences on protein function. Our findings highlight the importance of long-read sequencing for investigating differential alternative splicing usage and splicing factor function. These results have implications for understanding the role of *SF3B1* mutations in hematological malignancies and other diseases, and may be used in the future to predict new approaches for targeted therapies for these conditions.

Methods

Ethics approval

The study was approved by the Ethics Committee of the University of Cologne (Ethikvotum 11-319 from 11th December 2011, with an amendment from 7th June 2016) and the Ethics Committee of the University of Düsseldorf (Ethikvotum 3768, amendment from 24th October 2018). Informed consent has been obtained from all patients involved.

Cell lines and patients' samples

The isogenic cell line pairs, K562-SF3B1$^{K700E/wt}$ and its parental K562-SF3B1$^{wt/wt}$ (RRID:CVCL_0004), as well as Nalm6-SF3B1$^{H662Q/wt}$ and its parental Nalm6-SF3B1$^{wt/wt}$ were obtained from Horizon Discovery (HD181-012, HD115-110). Since homozygous *SF3B1* mutations were reported to be lethal (Lee et al. 2016), we used heterozygous cell lines. The K700E mutation is the most frequent *SF3B1* mutation reported in CLL and MDS, and H662Q mutation is also frequently reported (Wan and Wu 2013; Rossi et al. 2011; Quesada et al. 2012). The *SF3B1* mutated cell lines were described previously (Darman et al. 2015). HEK293-FT (RRID: CVCL_6911) was purchased from Thermo Fisher Scientific (#R70007). Information on cell line authentication and cell growth conditions are provided in Supplemental Methods.

CLL and B cell samples were obtained from the CLL-Biobank Cologne. *IGHV* mutational status was determined, as previously described (Rosenquist et al. 2017). Peripheral blood B cells were isolated

via negative selection using RosetteSep immunodensity-based cell separation (Stemcell Technologies, Vancouver, BC, Canada). The purity of CLL/ B cells was analyzed by flow cytometry and revealed that ≥ 90% cells co-expressed CD5/CD19.

Specimens from MDS with ring sideroblast (MDS-RS) patients were obtained from the MDS Biobank of the University Clinic Düsseldorf. Either RNA or cells were obtained from the Biobank. If cells were obtained, RNA was isolated using the Nucleospin RNA kit (Macerey Nagel). RNA quality was accessed by RNA ScreenTape analysis (Agilent) or a Bioanalyzer (Agilent).

Clinical information on the patients is summarized in Supplemental Table S1.

Plasmids

Plasmids pCMV-3Tag-1A-SF3B1^{wt} and pCMV-3Tag-1A-SF3B1^{K700E} (Alsafadi et al. 2016) were designed by Angelos Constantinou (Department of Molecular Bases of Human Diseases, CNRS UPR 1142, IGH-Institute of Human Genetics, Montpellier 34090, France) and kindly provided by Marc-Henri Stern, Institut Curie, Paris, France. Plasmids pcDNA3.1-FLAG-SF3B1-WT and pcDNA3.1-FLAG-hSF3B1-K700E (Kesarwani et al. 2017) were obtained from Addgene (#82576 and #82577). The human full-length *SF3B1* sequence has been previously reported to be impossible to clone into bacteria (Yokoi et al. 2011; Wang et al. 1998). Therefore, the plasmids consisted of synthetic sequences, codon-optimized for expression in bacteria (Alsafadi et al. 2016; Kesarwani et al. 2017). For the minigene constructs the intron and parts/complete adjacent upstream and downstream exons were PCR-amplified from K562 genomic DNA using Phusion Hot Start Flex polymerase (New England Biolabs) and cloned by the Hot Fusion (Fu et al. 2014) method into the BamHI restriction site of pcDNA3 (Invitrogen, https://www.addgene.org/vector-database/2092/). The open reading frame of the exons was left intact. The oligonucleotides used for cloning of the constructs are listed in Supplemental Table S5. Mutations and insertions were introduced by site directed mutagenesis using the Q5 site-directed mutagenesis kit (New England Biolabs). Oligonucleotides for site-directed mutagenesis were designed using the NEBaseChanger version 1.3.3 (New England Biolabs) and are

listed in Supplemental Table S5. All constructs were verified by Sanger sequencing (Microsynth Seqlab AG, Göttingen, Germany).

cDNA synthesis and validation of the splicing alterations

Transfections and RNA isolation were performed following standard procedures with PEI Max (PolyScience, #24765 1) and NucleoSpin RNA Mini kit (Macherey Nagel, #740955.250). An amount of 500 ng total RNA was reverse transcribed using SuperScript™ II (Thermo Fisher Scientific, #18064014) and hexamer oligonucleotides for the cDNA synthesis from K562 and Nalm6 RNA. For the minigene assays 500 ng RNA was reverse transcribed using SuperScript™ IV Reverse Transcriptase (Thermo Fisher Scientific, #18090010) with the plasmid-specific BGH-rev oligo in 20 µl. Subsequently, RNA in DNA-RNA hybrids was digested by RNase H incubation. For RT-PCR we used 1 µl of cDNA, Taq DNA polymerase, recombinant (Thermo Fisher Scientific, #10342020), and specific oligonucleotides (Supplemental Table S5) in a volume of 25 µl. The PCR ran for 35 PCR cycles. PCR products were separated on a 3–4% TAE-agarose gel.

PacBio Iso-Seq library preparation and sequencing

For the cDNA synthesis we used oligo(dT) oligonucleotides and the TeloPrime v2 Kit (Lexogen) to ensure the amplification of full-length mRNAs that contained a Cap-structure. Barcoded primers were used in the cDNA amplification step to enable multiplexing before library preparation. To enrich for slightly larger cDNAs, we adjusted the magnetic bead concentration in the bead clean-up after cDNA amplification. Subsequent library preparation was performed with the Express Template Kit 2.0 (PacBio).

In total, 58 libraries were sequenced on the PacBio Sequel II platform, by multiplexing 4 samples per 8M SMRT cell at the Genomics and Transcriptomics Laboratory, the production site of the West German Genome Center in Düsseldorf (Heinrich-Heine Universität, Düsseldorf, Germany) (Supplemental Table S1).

Processing of PacBio Iso-Seq data

Preprocessing of raw Iso-Seq sequencing data was performed with *Iso-Seq* (PacBio 2020 2021) software version 3.4, using the recommended parameters. In brief, we used the *ccs* tool to call circular consensus sequences by clustering and collapsing steps, *lima* to remove primers and adapters, and *isoseq refine* to demultiplex samples and filter out reads not featuring poly(A) sequences. This resulted more than 33 million aligned full-length non-chimeric (flnc) polyA HiFi reads (100,302 – 1,258,653 per library, average 582,135, Supplemental Table S1), with an average length of 2,721 bp. At this sequencing depth, transcript isoforms expressed at one transcript per million (TPM) are expected to be sequenced by at least 25 reads, and two TPM isoforms by at least 50 reads, with more than 95% probability (Supplemental Fig. S1).

According to the base quality values, 58 samples had an error rate of less than 1% in at least 99.7% of the HiFi reads and only four samples had higher error rates with 96.7% to 96.9% reads with < 1% error rate. Overall, the quality of the reads was high, with only 9% of reads potentially affected by technology-specific technical artefacts (Cocquet et al. 2006) (Supplemental Fig. S2).

The flnc reads were converted to FASTQ using SAMtools v.1.18 (Li and Durbin 2009), and, without an additional clustering step, aligned to the human genome GRCh38.p13 using minimap2 (Li 2018) version 2.22 with the preset parameters for high quality spliced reads (-ax splice:hq). For each sample, at least 99.85% of the reads were mapped to the genome, and at least 91.9% were uniquely mapped. Samples for which we sequenced more than one library were merged using SAMtools v.1.18 (Li and Durbin 2009) after the mapping step. Sequencing and mapping statistics per sample are detailed in Supplemental Table S1.

SF3B1 mutation calling was done with BCFtools v.1.13 (Danecek et al. 2021) mpileup and SnpEff v.5.1d (Cingolani et al. 2012).

Further analysis of Iso-Seq data was performed in Python v3, using IsoTools (Lienhard et al. 2023) version 0.2.8. In brief, aligned reads were imported and compared to the human reference annotation

version 36 from GENCODE (Frankish et al. 2021), to call, annotate, classify, and quantify transcripts, using IsoTools' add_sample_from_bam function.

For exploratory analysis, alternative splicing events were detected using IsoTools' alternative_splicing_events function. For each sample, individual events were quantified by percent spliced index (PSI) values, i.e., the number of reads supporting transcripts that include additional exonic sequence over all transcripts spanning that event. Based on these PSI values, principal component analysis (PCA) plots for different alternative splicing categories were computed using IsoTools' plot_embedding function.

Differential splicing events between *SF3B1*$^{mut/wt}$ and *SF3B1*$^{wt/wt}$ CLL, MDS and cell line samples were computed with the IsoTools altsplice_test function, using the betabinomial likelihood ratio test. This test models the variability within the tested groups with a beta-binomial mixture distribution, a binomial distribution where the probability parameter, *p*, of the binomial distribution *B(n,p)* follows a beta distribution, *Beta(a, b)*. The test compares the group-wise coverage of the splicing event with the total coverage.

$\Lambda = -2(l_0 - l_1)$, where:

$l_1 = \log\left(BB\left(k_1|\hat{\alpha}_1, \hat{\beta}_1, n_1\right)\right) + \log\left(BB\left(k_2|\hat{\alpha}_2, \hat{\beta}_2, n_2\right)\right)$ and

$l_0 = \log\left(BB\left(k_1 + k_2|\hat{\alpha}, \hat{\beta}, n_1 + n_2\right)\right)$.

Here, $BB(k|\alpha, \beta, n)$ is the probability mass function of the beta-binomial distribution and $\hat{\alpha}, \hat{\beta}$ are maximum likelihood estimates for the parameters. The maximum log-likelihood parameters are determined numerically by a quasi-Newton optimization method (LM-BFGS from SciPy (Virtanen et al. 2020)). Under the null hypothesis (i.e. no differential splicing) the test statistic is X^2 distributed with two degrees of freedom.

This formulation allows for considering within-group variability in a similar manner as tests based on negative binomial distribution for RNA-seq data, which is crucial for heterogeneous samples such as individual cancer patients. To be tested, we required the events to be covered by at least 10 reads in at

least 4 samples per group, and the minor alternative to be covered by at least 5% of the total reads (test = 'betabinom_lr', min_n = 10, min_sa = 4, min_alt_fraction = 0.05). Due to the limited number of samples, we did not include any covariates in the model analysis. We did not observe any bias towards highly expressed genes (Supplemental Fig. S41).

All 3' alternative splicing events (including those that were not differentially expressed between *SF3B1*$^{mut/wt}$ and *SF3B1*$^{wt/wt}$) were exported for further analysis (Supplemental Table S4).

Differential expression analysis on the Iso-Seq read counts was performed with DESeq2 (Love et al. 2014).

Estimation of branch point position and splice-site strength score

For all 3' alternative splicing events, we used the R (R Core Team 2021) Bioconductor package branchpointR (Signal et al. 2018) to predict branchpoint probabilities for both, canonical and alternative splice sites. We used the position with the highest branchpoint probability as the predicted branch point.

The strength of the 3' splice sites for the minigene constructs was calculated with SpliceRover (Zuallaert et al. 2018) (http://bioit2.irc.ugent.be/rover/splicerover, accessed on 23.06.2023 using the model "human acceptors").

Transcript coding potential

For a detailed description of the methods, please see Supplemental Notes.

Illumina RNA-seq library preparation and sequencing

RNA from K562 and Nalm6 cells was isolated using the NucleoSpin RNA Mini kit (Macherey Nagel) followed by DNase-digestion with the DNase I set (Zymo Research, #E1010) and clean-up using the NucleoSpin RNA Clean-up Mini Kit (Macherey-Nagel, #740948.50). RNA quality was surveyed using RNA ScreenTapes (Agilent) and the RNA integrity number (RIN) was 10 for all samples. CLL

cells from 19 CLL patients used for Iso-Seq and additional eight patients (including four patients with *SF3B1* mutation) were stored in RNA later and RNA was isolated using the RNeasy Mini Kit (Qiagen) followed by DNase digestion using the DNase I Amplification Grade Kit (Invitrogen) and a clean-up using RNeasy Min Elute columns (Qiagen). RNA-seq libraries for the cell lines and the CLL RNA were prepared using the TruSeq-stranded-total RNA SamplePrep kit (Illumina) according to the manufacturer's protocol. In brief, 2 µg of total RNA were depleted for ribosomal RNA using a Ribo-Zero-rRNA removal kit (Illumina), followed by random primed cDNA synthesis. Sequencing libraries were run at the sequencing core unit of the Max-Planck Institute for Molecular Genetics on a HiSeq 2500 (Illumina) using 50 bp paired-end reads.

Processing of Illumina RNA-seq data

We collected publicly available MDS RNA-seq data from the Short Read Archive using the following criteria: "MDS" in the study description, paired-end Illumina reads, available FASTQ files, human blood cell samples, and at least 5 samples per study. After merging technical replicates, our dataset consisted of 263 samples from *SF3B1*$^{wt/wt}$ MDS patients and 135 *SF3B1*$^{mut/wt}$ patients. All study and sample IDs are listed in Supplementary Table S1.

Illumina RNA-seq reads were aligned to the human reference genome GRCh38.p13 using STAR aligner version 2.7.6a (Dobin et al. 2013), with provided GFF annotation from GENCODE release 36 including annotation of non-chromosomal scaffolds. Alternative splicing events were called and quantified using rMATS (v4.1.1) (Shen et al. 2014).

Mutations in *SF3B1* were called as for Iso-Seq described above.

iCLIP experiments

iCLIP of K562-SF3B1$^{wt/wt}$ (three replicates) and K562-SF3B1$^{K700E/wt}$ (two replicates) was performed as described previously (Sutandy et al. 2016). To this end, exponentially growing K562 cells were pelleted, washed once with PBS and 10×10^6 were subjected to UV-cross linking at 400 mJ/cm^2 at

254 nm in 6 ml PBS in a 10 cm petri dish on ice. Crosslinked cells were scraped from the dish, collected by centrifugation for 2 min at 500 × g at 4 °C, snap frozen in liquid nitrogen and stored at -80 °C. About 3×10^6 cells were immunoprecipitated using 10 µg of a monoclonal SF3B1 antibody (clone 16, MBL D221-3) or 10 µg of the monoclonal U2AF2 antibody (U4759, SIGMA).

iCLIP data processing

Initial quality control was done using FastQC (Andrews 2010) before and after quality filtering). All reads having at least one position with a sequencing quality < 10 in the barcode area (positions 1–3, 4–7, 8–9) were removed. De-multiplexing and adapter trimming were done on quality filtered data using Flexbar (Roehr et al. 2017). No mismatches were allowed during de-multiplexing, while an error rate of 0.1 was accepted when trimming the adapter at the right end of the reads. Furthermore, a minimal overlap of 1 bp between reads and adapter was required and only trimmed reads with a minimal length of 15 bp (24 bp including the barcode) were kept for further analysis. Barcodes of remaining reads were trimmed (but kept as additional information with the reads). Trimmed reads were then mapped to genome assembly version GRCh38 using STAR (v. 2.6.1b) (Dobin et al. 2013) with 4% mismatched bases allowed and turned off soft-clipping on the 5-prime end, and GENCODE gene annotation v31 (Frankish et al. 2019). While technical duplicates were removed with UMI-tools (Smith et al. 2017) with *unique* method, all real duplicate reads were kept. Then, we checked the crosslink quality with iCLIPro (Hauer et al. 2015).

For facilitate comparisons, the crosslink events, i.e., reads after duplicate removal, of the replicates were randomly subsampled to the size of the smallest replicate (n = 13,892,358) (replicate 2 of SF3B1$^{K700E/wt}$; Supplemental Table S1).

Definition and classification of SF3B1 binding sites

Binding sites were identified from the merged crosslink events of SF3B $^{wt/wt}$ and SF3B1$^{K700E/wt}$ as described in Busch *et al.* (Busch et al. 2020) (Supplemental Fig. S29). For this, the crosslink events of all five replicates were combined and subjected to peak calling with PureCLIP (version 1.3.1) (Krakau et al. 2017) with default parameters. The PureCLIP-called sites (Psites) were filtered by first removing 5% of the Psites with the lowest score associated and then keeping only the top 20% of Psites within each gene annotated (GENCODE release 36, GRCh38; only annotations with a gene support level of 1 or 2 and transcript support level from 1 to 3). The Psites were then merged into

binding sites using the R/Bioconductor package BindingSiteFinder (version 1.0.3) (Brüggemann and Zarnack), using the following options: width of 5 nt (bsSize =5); ≥ 2 Psites (minWidth = 2, minClSites = 1) and ≥ 1 crosslink positions within each binding site (minCrosslinks = 1). In brief, Psites closer than 5 nt were merged into regions, and isolated Psites were discarded. Within each region, binding site centers were iteratively placed at the position with most crosslink events and extended by 2 nt on both sides. Binding site centers were required to harbor the maximum crosslink signal within the binding site. The optimal binding site width of 5 nt was determined by an evaluation of the ratio of crosslink events within binding sites of increasing width over the mean background signal in flanking windows of the same size (Supplemental Fig. S30). Next, binding sites that were not supported by all replicates in at least one condition ($SF3B1^{wt/wt}$ or $SF3B1^{K700E/wt}$) were filtered out (Supplemental Fig. S31). The threshold for sufficient coverage in a replicate was determined using the 5th percentile and a lower boundary of two crosslink events as described in Busch *et al.* (Busch et al. 2020). Finally, binding sites were assigned to target genes using GENCODE annotation (release 36, GRCh38; filtered as above) as described in Busch *et al.* (Busch et al. 2020). In total, this procedure identified 96,852 SF3B1 binding sites in 8,127 genes.

To classify distinct SF3B1 binding patterns in introns, bound regions were defined by merging intronic binding sites within a distance < 55 nt and resizing the obtained regions to 81 nt around the center, resulting in 56,224 regions harboring 87,199 binding sites. Following the approaches suggested in Heyl *et al*. (Heyl and Backofen 2021), we used unsupervised clustering to separate the crosslink patterns in the bound regions. For this, the crosslink coverage (sum of all replicates) was subjected to min-max normalization (Tarantola 2008) within each window (i.e., scaling such that the lowest and highest number of crosslink events are set to 0 and 1, respectively), followed by spline-smoothing using the smooth.spline function (R package stats, version 4.1.0) with lambda 0.2 (spar = 0.2) and inflated dimensions (dim = 150). This changed the shape of the matrix from A × B (56,224 × 81) to A × B' (56,244 × 150), where A is the number of regions, B are the nucleotide positions and B' are the inflated nucleotide positions. The matrix A × B' of normalized and smoothed crosslink coverages was then subjected to dimension reduction using uniform manifold approximation and

projection (UMAP) (McInnes et al. 2018) with the *umap* function (package umap, version 0.2.7) with parameters n_epochs = 5000, n_components = 2, min_dist = 0.01 and n_neighbors = 5 (Supplemental Fig. S34A). The UMAP results were assigned to clusters using density-based clustering of applications with noise (DBSCAN) (Ester et al. 1996) with the dbscan function (R package dbscan (Hahsler et al. 2019), version 1.1, eps = 0.3), with a minimum number of 150 points per cluster (MinPts = 150), yielding three clusters: C1 (n = 35,907 regions), C2 (n = 5,635) and C3 (n = 13,847). Bound regions in cluster C0 (n = 835) were deemed as outliers that could not be assigned to any of the fitted density centers and excluded from further analysis. Bound regions in cluster C3 (wide pattern) were smoothed more finely (spar = 0.1, dim = 500) and then subjected to a second round of UMAP dimension reduction (parameters as above) and DBSCAN clustering (MinPts = 60, eps = 0.23), yielding subclusters #0–33 (Supplemental Fig. S34B). Cluster numbering is based on the increasing distance between the two modes in the arrangement of binding sites, calculated on the summed and smoothed coverages within each cluster using the locmodes function (R package multimode, version 1.5 (Ameijeiras-Alonso et al. 2021)) (Supplemental Fig. S34C).

Data Access

Cell lines' and patients' transcriptome raw FASTQ files or PacBio CCS unaligned BAM files have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1037338 or to the European Genome-Phenome Archive (https://ega-archive.org/) under accession number EGAS50000000053, respectively. iCLIP raw FASTQ data and processed files have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE247658. Patient data access will be controlled by the Data Access Committee at the Institute for Translational Epigenetics at the University Hospital Cologne, University of Cologne, Cologne, Germany.

The complete Iso-Seq/iCLIP data analysis code is available as a Jupyter Notebook in the GitHub repository https://github.com/ZarnackGroup/go_long2023 as well at Zenodo at the following link https://zenodo.org/doi/10.5281/zenodo.12597945.

Competing Interest Statement

MH: Honoraria: (speaker´s bureau and/or advisory board): Roche, Janssen, Abbvie; Research support: Roche, Janssen, Abbvie, Astra Zeneca, Beigene.

Acknowledgements

The study was funded by the German Research Foundation: KFO286-RP8/SCHW1605/1-1, SCHW1605/4-1 (GO-LONG), SFB1399 and SFB1530 to M.R.S., KFO-286-RP6 to M. H., KFO-286-CP to C.D.H., SFB1530 to M.H., the Volkswagen Stiftung Lichtenberg program to M.R.S. and the Center for Molecular Medicine Cologne, CMMC (A12 to M.R.S.).

We acknowledge IMB Genomics Core Facility and its NextSeq 500 sequencer [funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) INST 247/870-1 FUGG]. We would like to thank Angelos Constantinou (IGH-Institute of Human Genetics, France) and Marc-Henri Stern (Institut Curie, Paris, France) for sharing the *SF3B1* plasmids, Anke Busch (IMB Mainz, Germany) for iCLIP data pre-processing, Elena Wasserburger-Zichel (University Hospital Cologne, Germany), and Bernd Timmermann (Sequencing Core Unit, Max Planck Institute for Molecular Genetics, Berlin, Germany) for their technical assistance. We furthermore thank the Regional Computing Center of the University of Cologne (RRZK) for providing computing time on the DFG-funded (Funding number: INST 216/512/1FUGG) High Performance Computing (HPC) system CHEOPS as well as IT support. In addition, we acknowledge computational support of the Centre for Information and Media Technology, especially the HPC team at the Heinrich-Heine University, as well as the computing time provided by the John von Neumann Institute for Computing on the

supercomputer JUWELS at Jülich Supercomputing Centre (user IDs: VSK33). This work was supported by the DFG Research Infrastructure West German Genome Center (407493903) as part of the Next Generation Sequencing Competence Network (project 423957469). High-throughput sequencing was carried out at the West German Genome Center, and production sites in Cologne and Düsseldorf.

Authors' contributions

M.R.S., C.G., R.H., K.Z., H.G., N.G., M.H., and J.K. designed the study. H.H., L.B., A.K., K.B., K.K. and C.D.H. acquired the data. A.P., M.L., M.B., J.K., H.G., C.G., R.H., K.Z., and M.R.S. analyzed and interpreted the data. A.P., M.L., C.G., R.H., K.Z., and M.R.S. drafted and wrote the manuscript. All authors have read and approved the final manuscript.

Figure legends

Figure 1. Long-read sequencing of CLL and MDS patient samples discovers novel isoforms. (*A*) Distribution of *SF3B1* mutations in CLL and MDS patient samples used for Iso-Seq: each dot represents a mutated sample. One CLL patient is marked twice due to two mutations (I704N and D894G). Note that the A284T mutation is outside the HEAT repeat domains, and thus was grouped as a wild-type sample, also according to further analysis described below. SF3B1 is shown as the major isoform expressed, with the full splice match to annotated isoform 201. **(*B*)** The number of substantial transcripts identified in each group and expressed at the level of at least 1 transcript per million (TPM) colored by the category of isoform novelty: full splice match (FSM), with incomplete splice matches (ISM), with combinations of annotated splice junctions (novel in catalog, NIC), with at least one novel splice site (novel not in catalog, NNC), or from novel genes (NOVEL) (Tardaguila et al. 2018). **(*C*)** Venn diagrams showing the overlap between isoforms from (*B*) expressed at ≥ 1 TPM in each group. **(*D*)** *SF3B1* isoforms expressed at > 10% relative expression level. **(*E*)** Relative expression levels of *SF3B1* isoforms from (*D*).

Figure 2. *SF3B1* mutation effect is independent of the biological background, but its manifestation depends on the transcriptomic profile. (*A*) The number of alternative splicing events identified with Iso-Seq separated by splicing event type in all groups investigated, differentiated by the novelty class. **(*B*)** Overlap between significantly altered alternative splicing events (ASEs) in samples with *SF3B1* mutation identified in the three datasets used (cell lines, CLL patients, or MDS patients). **(*C*)** Correlation of isoform usage measured by the difference in PSI from all events listed in (*B*) i.e., events called significant in at least one of the three datasets are shown. The colors of the dots correspond to significance reached only in one set (blue: X axis only; red: Y axis only; light blue: both; grey: none, i.e. called significant in a dataset absent from the graph). Pearson correlation coefficient (*R*) and associated p-value (*p*) are given. **(*D*)** Violin plots with boxplots show the distribution of expression values of the genes with dataset-specific ASEs from (*C*). Significant differences are marked with *** for paired, two-tailed Student's *t*-test p-value < 0.001, ** p-value < 0.01, * p-value < 0.05, N.S. – not significant with p-value ≥ 0.05.

Figure 3. *SF3B1* mutation increases 3' alternative splice sites usage and decreases intron retention. (*A*) Number of differential alternative splicing events with shorter or longer variant expression (percent spliced index, PSI) in $SF3B1^{mut/wt}$ vs. $SF3B1^{wt/wt}$ samples. **(*B*)** Highly significantly altered (q-value < 0.01) 3' alternative splice sites clearly separate samples by *SF3B1* mutations in leukemia cell lines as well as CLL and MDS patients based on the longer variant PSI values showing four clusters described more in Supplementary Fig. S19. **(*C*)** Principal component (PC) analysis, based on the isoform usage of 3' alternative splice sites, clearly separates CLL and MDS patients, as well as cell lines, according to the *SF3B1* mutational status. **(*D*)** Swarm plots showing the distribution of the isoform usage (PSI) among groups with or without *SF3B1* mutation. **(*E*)** Validation of the differential splicing associated with *SF3B1* mutation with RT-PCR experiment in isogenic K562 and Nalm6 cell lines. **(*F*)** Minigene assays workflow. HEK293T cells were co-transfected with minigenes and either $SF3B1^{wt}$ or $SF3B1^{K700E}$ for 48 h. RNA was extracted and used for amplification of splicing products with minigene-specific primers. **(*G*)** The results from the minigene assay from (*F*). The

lower band in the agarose gel corresponds to the usage of the canonical AG, the upper band to the upstream AG'.

Figure 4. *SF3B1* mutation results in altered mRNAs potentially translated into modified proteins. (*A*) Coding potential of the known and novel isoforms identified divided by CDS similarity to annotated isoforms and NMD prediction. **(*B*)** Effect of *SF3B1*^{mut}-associated ASEs on the protein-coding potential. **(*C*)** *SF3B1* isoforms detected in this study with CLL- or MDS-specific isoforms. **(*D*)** The *PRPF38A* isoform expression levels (left) and structure with Pfam domains indicated (right). Highlighted in yellow is the *SF3B1*^{mut}-associated 3'AS that may influence the protein function. **(*E*)** Schematic of major *MAP3K7* isoforms (left) with the protein kinase domain showed as green boxes. The *SF3B1*^{mut}-associated 3'AS is highlighted in yellow and ORF start/end indicated by the triangles. Highlighted in light green are predicted upstream ORFs (uORFs). Red box highlights the additional exon 12 in the isoform 202, which is absent in the isoform 205. Expression of each isoform is shown on the right. The expression of isoforms with 3'AS event is shown as striped bars.

Figure 5. *SF3B1* mutations promote upstream alternative 3' splicing sites and partially dependent on the sequence context. (*A*) 3' alternative splice site distance distribution. Negative distances indicate an alternative was located upstream and positive values indicate alternative located downstream leading to a shorter exon. Blue line represents proportion and green total number of 3'alternative splicing events (3'AS). Dotted vertical lines indicate the enriched region of 12–21 nt upstream of the canonical AG. **(*B*)** Minigene assays with long (45–50 bases) AG–AG' inserts, shortened inserts containing about 20 nt directly upstream the AG including the polypyrimidine (Py) tract, and short (15–20 nt) AG–AG' inserts from non-differentially alternatively spliced 3'AS events. The chosen events without differential splicing detected with *SF3B1* mutation were from *PAPCL1*, *USP1* and *ZNF124* (AG'–AG distance > 50 nt) as well as *GPR98B*, *UROD* and *CELF2* (AG'–AG distance < 20 nt). **(*C*)** Table showing splice site strength for AG and AG' calculated with SpliceRover (Zuallaert et al., 2018).

Figure 6. Multimodal SF3B1 binding. (*A*) SF3B1 choice of AG within a narrow window of 12–21 nt is strongly affected by the alternative splice site distance. The AG usage is shown as percent spliced index (PSI) as a function of the splice site distance in SF3B1$^{wt/wt}$. Rolling mean across 20 nt and smoothed (loess method) trend line are shown for upstream (violet) and downstream (pink) AG. **(*B*)** Schematic workflow of processing iCLIP reads and calling SF3B1 binding sites. **(*C*)** Defining optimal site binding width. A binding site width of 5 nt optimally captures the SF3B1 crosslink events. Dot plot shows average ratio of crosslink events within binding sites of increasing widths (x-axis) over the mean background signal in flanking windows of the same size, indicating how much more signal occurs within the binding sites compared to their immediate surrounding. **(*D*)** SF3B1 binding sites reproducibility across replicates. Upper panel shows overlaps of supported binding sites in the replicates, with threshold for sufficient coverage individually adjusted to the signal depth in each replicate (Busch et al. 2020)**. (*E*)** Gene classes targeted by SF3B1 based on iCLIP. **(*F*)** Transcript regions of protein-coding genes targeted by the SF3B1 based on iCLIP. **(*G*)** Differential SF3B1 binding sites in K562-SF3B1$^{K700E/wt}$ vs. K562-SF3B1$^{wt/wt}$. **(*H*)** Meta-profiles of SF3B1 (top) and U2AF2 (bottom) binding centered at branch point adenosine (left) and 3' splice site AG (right). **(*I*)** Distribution of distances between neighboring binding sites. **(*J*)** Density plot and heat map showing SF3B1 patterns in regions from three cluster types identified in Supplemental Fig. S35. **(*K*)** Distribution of the number of binding sites per region for each of the three clusters. **(*L*)** Distribution of the distance between SF3B1 binding region and closest splice site.

Figure 7. SF3B1 promotes alternative proximal AG' usage. (*A*) U2AF2 and SF3B1 binding to pre-mRNA based on the iCLIP signal. U2AF2 iCLIP was performed using K562-SF3B1$^{wt/wt}$ cells and SF3B1 iCLIP was performed on K562-SF3B1$^{wt/wt}$ and K562-SF3B1$^{K700E/wt}$. On the left panel the splice site distance is shown, followed by the iCLIP signal aligned to the more downstream AG used. Significance of the differential 3'AS usage between K562-SF3B1$^{wt/wt}$ and K562-SF3B1$^{K700E/wt}$ is shown as annotation bar on the right side of the iCLIP signal heatmap. **(*B*)** For every 3'AS from (*A*) the upstream AG PSI value is denoted for *SF3B1*$^{wt/wt}$ (blue, left), *SF3B1*$^{mut/wt}$ (red, center), and the difference between *SF3B1*$^{mut/wt}$ and *SF3B1*$^{wt/wt}$ (black, right). **(*C*)** Distribution of AG occurrence (top),

proportion of significantly alternatively used AG' (middle) and AG' scores among introns with significant (pink) or non-significant (violet) difference in usage between *SF3B1*$^{mut/wt}$ and *SF3B1*$^{wt/wt}$. **(*D*)** Number of alternative AGs (AG') among regions with multiple AG's. AG's within 6 nt distance from AG were removed to avoid NAGNAG acceptor sites (Hiller et al. 2004). **(*E*)** Distance of AG's and BPs from AG among non-significant (grey) and significant (red) differentially splicing events. **(*F*)** Splice site strength of canonical and alternative AGs as calculated with MaxEnt score (Yeo and Burge 2003). For figures (*A*-*F*) only 3'ASs with the following features were used: i) placed chromosome scaffold; ii) classical AGs; iii) an intron became shorter in the mutant (the use of upstream alternative AG'); and iv) overlap with an iCLIP crosslink. The more used AG in *SF3B1*$^{wt/wt}$ was set as canonical AG.

References

Agafonov DE, Deckert J, Wolf E, Odenwälder P, Bessonov S, Will CL, Urlaub H, Lührmann R. 2011. Semiquantitative Proteomic Analysis of the Human Spliceosome via a Novel Two-Dimensional Gel Electrophoresis Method. *Mol Cell Biol* **31**: 2667–2682

Alsafadi S, Dayot S, Tarin M, Houy A, Bellanger D, Cornella M, Wassef M, Waterfall JJ, Lehnert E, Roman-Roman S, et al. 2021. Genetic alterations of SUGP1 mimic mutant-SF3B1 splice pattern in lung adenocarcinoma and other cancers. *Oncogene* **40**: 85–96.

Alsafadi S, Houy A, Battistella A, Popova T, Wassef M, Henry E, Tirode F, Constantinou A, Piperno-Neumann S, Roman-Roman S, et al. 2016. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. *Nat Commun* **7**: 10615.

Ameijeiras-Alonso J, Crujeiras RM, Rodriguez-Casal A. 2021. multimode: An R Package for Mode Assessment. *J Stat Softw* **97**: 1–32.

Andrews S. 2010. FASTQC - a quality control tool for high throughput sequence data. www.bioinformatics.babraham.ac.uk/projects/fastqc.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* **57**: 289–300.

Bland P, Saville H, Wai PT, Curnow L, Muirhead G, Nieminuszczy J, Ravindran N, John MB, Hedayat S, Barker HE, et al. 2023. SF3B1 hotspot mutations confer sensitivity to PARP inhibition by eliciting a defective replication stress response. *Nat Genet* **55**: 1311–1323.

Bradley RK, Anczuków O. 2023. RNA splicing dysregulation and the hallmarks of cancer. *Nat Rev Cancer* **23**: 135–155.

Brüggemann M, Zarnack K. Binding site defintion based on iCLIP data. https://bioconductor.org/packages/release/bioc/html/BindingSiteFinder.html.

Busch A, Brüggemann M, Ebersberger S, Zarnack K. 2020. iCLIP data analysis: A complete pipeline from sequencing reads to RBP binding sites. *Methods* **178**: 49–62.

Canbezdi C, Tarin M, Houy A, Bellanger D, Popova T, Stern M-H, Roman-Roman S, Alsafadi S. 2021. Functional and conformational impact of cancer-associated SF3B1 mutations depends on the position and the charge of amino acid substitution. *Comput Struct Biotechnol J* **19**: 1361–1370.

Carrocci TJ, Zoerner DM, Paulson JC, Hoskins AA. 2017. SF3b1 mutations associated with myelodysplastic syndromes alter the fidelity of branchsite selection in yeast. *Nucleic Acids Res* **45**: 4837–4852.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**: 80–92.

Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88**: 127–131.

Cortés-López M, Chamely P, Hawkins AG, Stanley RF, Swett AD, Ganesan S, Mouhieddine TH, Dai X, Kluegel L, Chen C, et al. 2023. Single-cell multi-omics defines the cell-type-specific impact of splicing aberrations in human hematopoietic clonal outgrowths. *Cell Stem Cell*.

Cretu C, Schmitzová J, Ponce-Salvatierra A, Dybkov O, De Laurentiis EI, Sharma K, Will CL, Urlaub H, Lührmann R, Pena V. 2016. Molecular Architecture of SF3b and Structural Consequences of Its Cancer-Related Mutations. *Mol Cell* **64**: 307–319.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**: giab008.

Darman RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, Bailey SL, Bhavsar EB, Chan B, Colla S, et al. 2015. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. *Cell Rep* **13**: 1033–1045.

DeBoever C, Ghia EM, Shepard PJ, Rassenti L, Barrett CL, Jepsen K, Jamieson CHM, Carson D, Kipps TJ, Frazer KA. 2015. Transcriptome sequencing reveals potential mechanism of cryptic 3' splice site selection in SF3B1-mutated cancers. *PLoS Comput Biol* **11**: e1004105.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Dolatshad H, Pellagatti A, Liberante FG, Llorian M, Repapi E, Steeples V, Roy S, Scifo L, Armstrong RN, Shaw J, et al. 2016. Cryptic splicing events in the iron transporter ABCB7 and other key target genes in SF3B1-mutant myelodysplastic syndromes. *Leukemia* **30**: 2322–2331.

Ester M, Kriegel H-P, Sander J, Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. pp. 226–231, AAAI Press.

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**: D766–D773.

Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, et al. 2021. GENCODE 2021. *Nucleic Acids Res* **49**: D916–D923.

Fu C, Donovan WP, Shikapwashya-Hasser O, Ye X, Cole RH. 2014. Hot Fusion: an efficient method to clone multiple DNA fragments as well as inverted repeats without ligase. *PLoS One* **9**: e115318.

Gozani O, Feld R, Reed R. 1996. Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes Dev* **10**: 233–243.

Gozani O, Potashkin J, Reed R. 1998. A Potential Role for U2AF-SAP 155 Interactions in Recruiting U2 snRNP to the Branch Site. *Mol Cell Biol* **18**: 4752–4760.

Hahsler M, Piekenbrock M, Doran D. 2019. dbscan: Fast Density-Based Clustering with R. *J Stat Softw* **91**: 1–30.

Hauer C, Curk T, Anders S, Schwarzl T, Alleaume A-M, Sieber J, Hollerer I, Bhuvanagiri M, Huber W, Hentze MW, et al. 2015. Improved binding site assignment by high-resolution mapping of RNA–protein interactions using iCLIP. *Nat Commun* **6**: 7921.

Heyl F, Backofen R. 2021. StoatyDive: Evaluation and classification of peak profiles for sequencing data. *Gigascience* **10**: giab045.

Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, Schreiber S, Backofen R, Platzer M. 2004. Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nat Genet* **36**: 1255–1257.

Hooper JE. 2014. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Hum Genomics* **8**: 3.

Huang L, Lou C-H, Chan W, Shum EY, Shao A, Stone E, Karam R, Song H-W, Wilkinson MF. 2011. RNA Homeostasis Governed by Cell Type-Specific and Branched Feedback Loops Acting on NMD. *Mol Cell* **43**: 950–961.

Inoue D, Chew G-L, Liu B, Michel BC, Pangallo J, D'Avino AR, Hitchman T, North K, Lee SC-W, Bitner L, et al. 2019. Spliceosomal disruption of the non-canonical BAF complex in cancer. *Nature* **574**: 432–436.

Kesarwani AK, Ramirez O, Gupta AK, Yang X, Murthy T, Minella AC, Pillai MM. 2017. Cancer-associated SF3B1 mutants recognize otherwise inaccessible cryptic 3′ splice sites within RNA secondary structures. *Oncogene* **36**: 1123–1133.

König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J. 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**: 909–915.

Krakau S, Richard H, Marsico A. 2017. PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biol* **18**: 240.

Landau DA, Tausch E, Taylor-Weiner AN, Stewart C, Reiter JG, Bahlo J, Kluth S, Bozic I, Lawrence M, Böttcher S, et al. 2015. Mutations driving CLL and their evolution in progression and relapse. *Nature*.

Lee SC-W, Dvinge H, Kim E, Cho H, Micol J-B, Chung YR, Durham BH, Yoshimi A, Kim YJ, Thomas M, et al. 2016. Modulation of splicing catalysis for therapeutic targeting of leukemia with mutations in genes encoding spliceosomal proteins. *Nat Med* **22**: 672–678.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Lienhard M, van den Beucken T, Timmermann B, Hochradel M, Börno S, Caiment F, Vingron M, Herwig R. 2023. IsoTools: a flexible workflow for long-read transcriptome sequencing analysis. *Bioinformatics* **39**: btad364.

Liu Z, Yoshimi A, Wang J, Cho H, Chun-Wei Lee S, Ki M, Bitner L, Chu T, Shah H, Liu B, et al. 2020. Mutations in the RNA Splicing Factor SF3B1 Promote Tumorigenesis through MYC Stabilization. *Cancer Discov* **10**: 806–821.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Malcovati L, Karimi M, Papaemmanuil E, Ambaglio I, Jädersten M, Jansson M, Elena C, Gallì A, Walldin G, Della Porta MG, et al. 2015. SF3B1 mutation identifies a distinct subset of myelodysplastic syndrome with ring sideroblasts. *Blood* **126**: 233–241.

Malcovati L, Stevenson K, Papaemmanuil E, Neuberg D, Bejar R, Boultwood J, Bowen DT, Campbell PJ, Ebert BL, Fenaux P, et al. 2020. SF3B1-mutant MDS as a distinct disease subtype: a proposal from the International Working Group for the Prognosis of MDS. *Blood* **136**: 157–170.

McInnes L, Healy J, Melville J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* 1802.03426.

Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. 2015. Genome-wide discovery of human splicing branchpoints. *Genome Res* **25**: 290–303.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**: D412–D419.

Obeng EA, Chappell RJ, Seiler M, Chen MC, Campagna DR, Schmidt PJ, Schneider RK, Lord AM, Wang L, Gambe RG, et al. 2016. Physiologic Expression of Sf3b1(K700E) Causes Impaired Erythropoiesis, Aberrant Splicing, and Sensitivity to Therapeutic Spliceosome Modulation. *Cancer Cell* **30**: 404–417.

PacBio 2020. 2021. IsoSeq v3: Scalable De Novo Isoform Discovery [Online]. https://github.com/PacificBiosciences/IsoSeq.

Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C, et al. 2011. Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts. *New England Journal of Medicine* **365**: 1384–1395.

Pellagatti A, Armstrong RN, Steeples V, Sharma E, Repapi E, Singh S, Sanchi A, Radujkovic A, Horn P, Dolatshad H, et al. 2018. Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. *Blood* **132**: 1225–1240.

Peng H, Jabbari JS, Tian L, Chua CC, Anstee NS, Amin N, Wei AH, Davidson NM, Roberts AW, Huang DCS, et al. 2024. Single-cell Rapid Capture Hybridization sequencing (scRaCH-seq) to reliably detect isoform usage and coding mutations in targeted genes at a single-cell level. *bioRxiv* 2024.01.30.577942.

Pineda JMB, Bradley RK. 2018. Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev* **32**: 577–591.

Porter DF, Miao W, Yang X, Goda GA, Ji AL, Donohue LKH, Aleman MM, Dominguez D, Khavari PA. 2021. easyCLIP analysis of RNA-protein interactions incorporating absolute quantification. *Nat Commun* **12**: 1569.

Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, Ramsay AJ, Beà S, Pinyol M, Martínez-Trillos A, et al. 2012. Exome sequencing identifies recurrent mutations of the splicing factor *SF3B1* gene in chronic lymphocytic leukemia. *Nat Genet*.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria* **0**: {ISBN} 3-900051-07-0.

Rehrauer H, Opitz L, Tan G, Sieverling L, Schlapbach R. 2013. Blind spots of quantitative RNA-seq: the limits for assessing abundance, differential expression, and isoform switching. *BMC Bioinformatics* **14**: 370.

Roehr JT, Dieterich C, Reinert K. 2017. Flexbar 3.0 – SIMD and multicore parallelization. *Bioinformatics* **33**: 2941–2942.

Rosenquist R, Ghia P, Hadzidimitriou A, Sutton LA, Agathangelidis A, Baliakas P, Darzentas N, Giudicelli V, Lefranc MP, Langerak AW, et al. 2017. Immunoglobulin gene sequence analysis in chronic lymphocytic leukemia: Updated ERIC recommendations. *Leukemia*.

Rossi D, Bruscaggin A, Spina V, Rasi S, Khiabanian H, Messina M, Fangazio M, Vaisitti T, Monti S, Chiaretti S, et al. 2011. Mutations of the *SF3B1* splicing factor in chronic lymphocytic leukemia: Association with progression and fludarabine-refractoriness. *Blood*.

Schmitzová J, Cretu C, Dienemann C, Urlaub H, Pena V. 2023. Structural basis of catalytic activation in human splicing. *Nature* **617**: 842–850.

Seiler M, Peng S, Agrawal AA, Palacino J, Teng T, Zhu P, Smith PG, Buonamici S, Yu L. 2018. Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. *Cell Rep* **23**: 282-296.e4.

Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. 2014. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences* **111**: E5593 LP-E5601.

Shiozawa Y, Malcovati L, Gallì A, Sato-Otsubo A, Kataoka K, Sato Y, Watatani Y, Suzuki H, Yoshizato T, Yoshida K, et al. 2018. Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. *Nat Commun*.

Signal B, Gloss BS, Dinger ME, Mercer TR. 2018. Machine learning annotation of human branchpoints. *Bioinformatics* **34**: 920–927.

Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**: 491–499.

Sutandy FXR, Hildebrandt A, König J. 2016. Profiling the Binding Sites of RNA-Binding Proteins with Nucleotide Resolution Using iCLIP. In *Post-Transcriptional Gene Regulation* (ed. E. Dassi), pp. 175–195, Springer New York, New York, NY.

Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun*.

Tang Q, Rodriguez-Santiago S, Wang J, Pu J, Yuste A, Gupta V, Moldón A, Xu Y-Z, Query CC. 2016. SF3B1/Hsh155 HEAT motif mutations affect interaction with the spliceosomal ATPase Prp5, resulting in altered branch site selectivity in pre-mRNA splicing. *Genes Dev* **30**: 2710–2723.

Tarantola S. 2008. European innovation scoreboard: Strategies to measure country progress over time. *JRC scientific and Technical Reports, EUR* **23526**.

Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al. 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.*

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**: 261–272.

Wan Y, Wu CJ. 2013. SF3B1 mutations in chronic lymphocytic leukemia. *Blood*.

Wang C, Chua K, Seghezzi W, Lees E, Gozani O, Reed R. 1998. Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis. *Genes Dev* **12**: 1409–1414.

Wang L, Brooks AN, Fan J, Wan Y, Gambe R, Li S, Hergert S, Yin S, Freeman SS, Levin JZ, et al. 2016. Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia. *Cancer Cell* **30**: 750–763.

Yang F, Bian T, Zhan X, Chen Z, Xing Z, Larsen NA, Zhang X, Shi Y. 2023. Mechanisms of the RNA helicases DDX42 and DDX46 in human U2 snRNP assembly. *Nat Commun* **14**: 897.

Yang H, Beutler B, Zhang D. 2021. Emerging roles of spliceosome in cancer and immunity. *Protein Cell*.

Yeo G, Burge CB. 2003. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, *RECOMB '03*, pp. 322–331, Association for Computing Machinery, New York, NY, USA.

Yokoi A, Kotake Y, Takahashi K, Kadowaki T, Matsumoto Y, Minoshima Y, Sugi NH, Sagane K, Hamaguchi M, Iwata M, et al. 2011. Biological validation that SF3b is a target of the antitumor macrolide pladienolide. *FEBS J* **278**: 4870–4880.

Zarnack K, König J, Tajnik M, Martincorena I, Eustermann S, Stévant I, Reyes A, Anders S, Luscombe NM, Ule J. 2013. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**: 453–466.

Zhang C, Zhang B, Lin L-L, Zhao S. 2017. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18**: 583.

Zhang J, Ali AM, Lieu YK, Liu Z, Gao J, Rabadan R, Raza A, Mukherjee S, Manley JL. 2019. Disease-Causing Mutations in SF3B1 Alter Splicing by Disrupting Interaction with SUGP1. *Mol Cell* **76**: 82-95.e7.

Zhang J, Huang J, Xu K, Xing P, Huang Y, Liu Z, Tong L, Manley JL. 2022. DHX15 is involved in SUGP1-mediated RNA missplicing by mutant SF3B1 in cancer. *Proceedings of the National Academy of Sciences* **119**: e2216712119.

Zhang J, Xie J, Huang J, Liu X, Xu R, Tholen J, Galej WP, Tong L, Manley JL, Liu Z. 2023. Characterization of the SF3B1–SUGP1 interface reveals how numerous cancer mutations cause mRNA missplicing. *Genes Dev*.

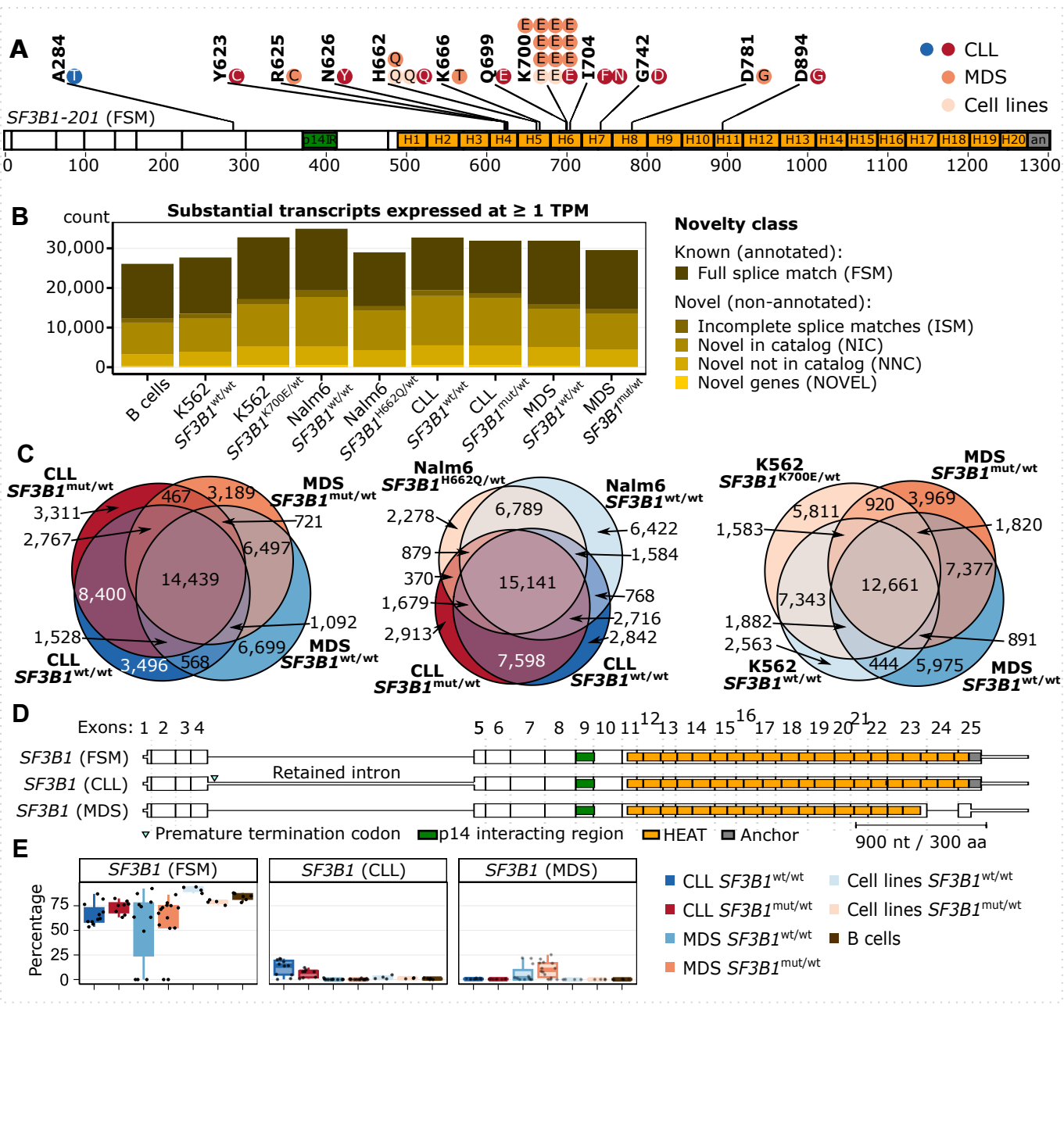
Zhang X, Zhan X, Bian T, Yang F, Li P, Lu Y, Xing Z, Fan R, Zhang QC, Shi Y. 2024. Structural insights into branch site proofreading by human spliceosome. *Nat Struct Mol Biol* **31**: 835–845.

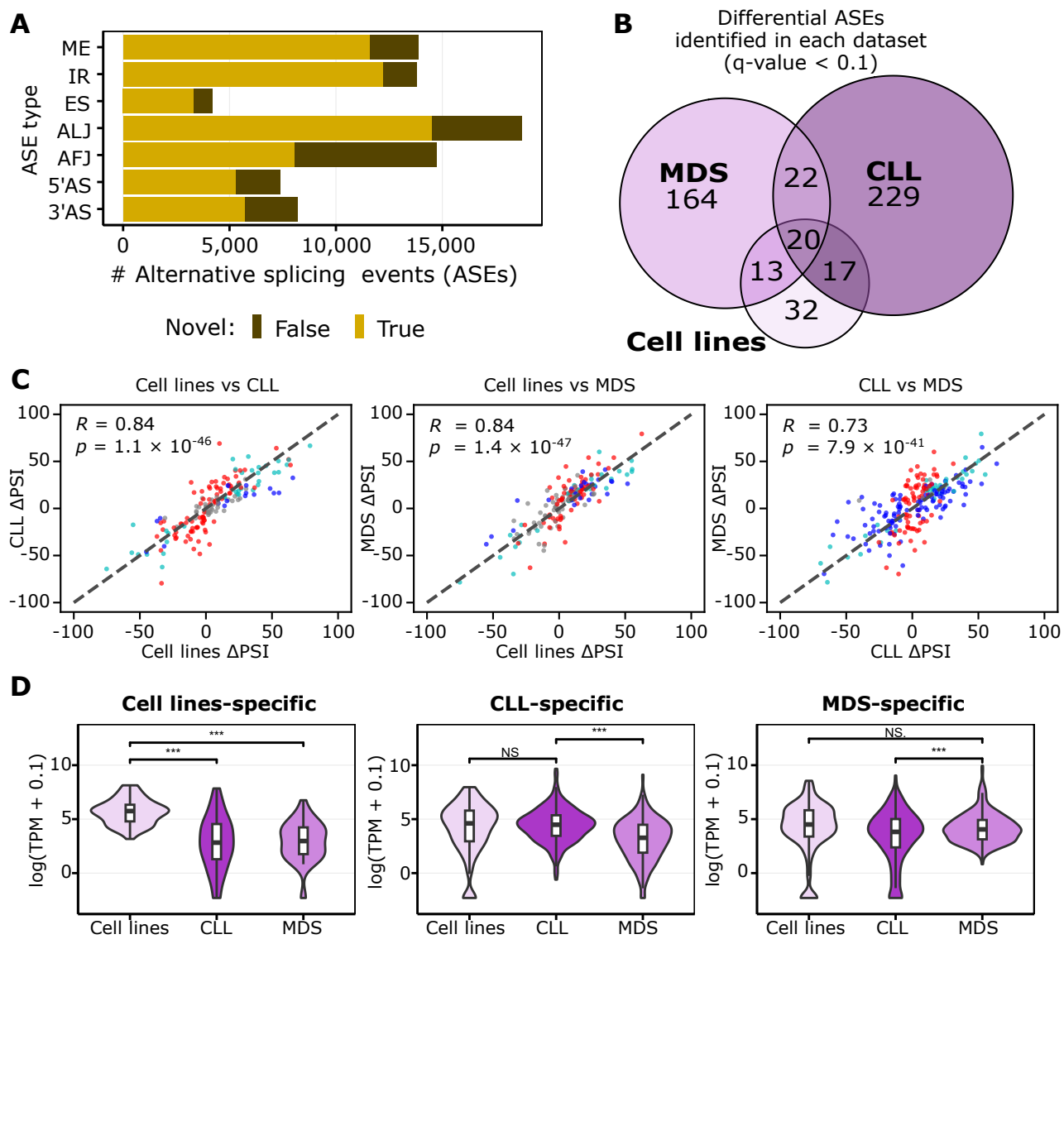
Zhang Z, Rigo N, Dybkov O, Fourmann J-B, Will CL, Kumar V, Urlaub H, Stark H, Lührmann R. 2021. Structural insights into how Prp5 proofreads the pre-mRNA branch site. *Nature* **596**: 296–300.

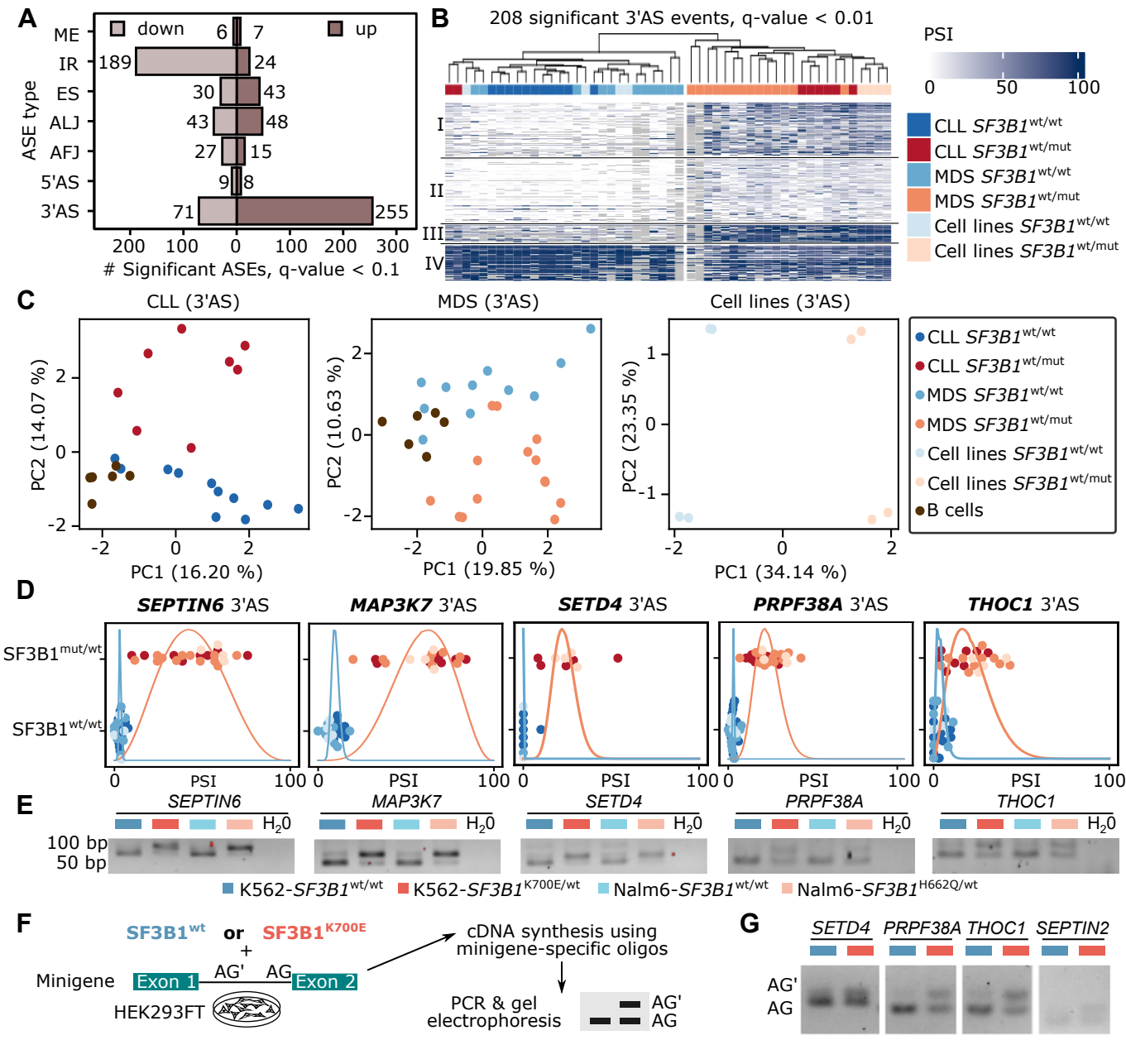
Zhao B, Li Z, Qian R, Liu G, Fan M, Liang Z, Hu X, Wan Y. 2022. Cancer-associated mutations in SF3B1 disrupt the interaction between SF3B1 and DDX42. *The Journal of Biochemistry* **172**: 117–126.

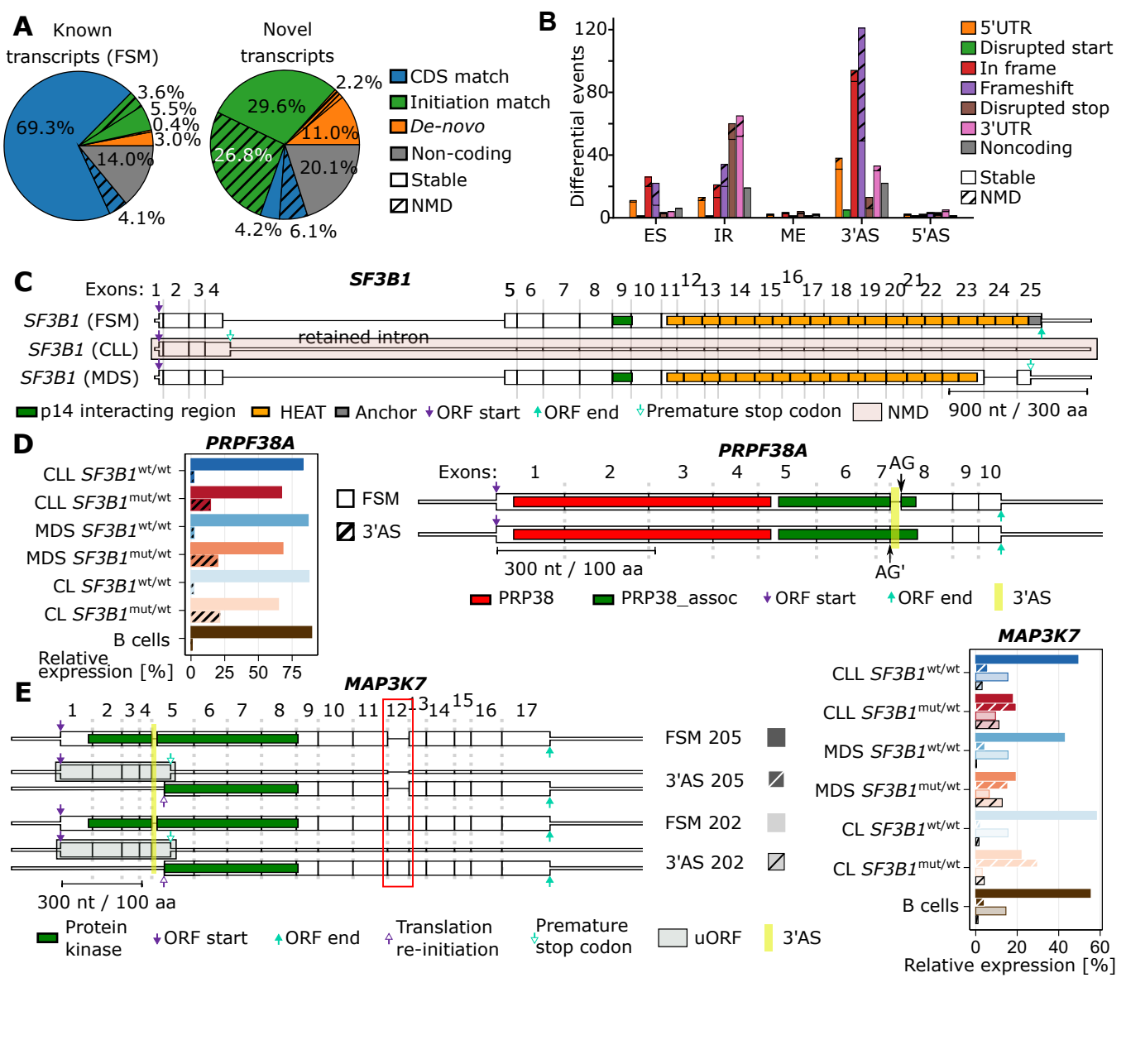
Zhou Z, Gong Q, Wang Y, Li M, Wang L, Ding H, Li P. 2020. The biological function and clinical significance of SF3B1 mutations in cancer. *Biomark Res* **8**: 38.

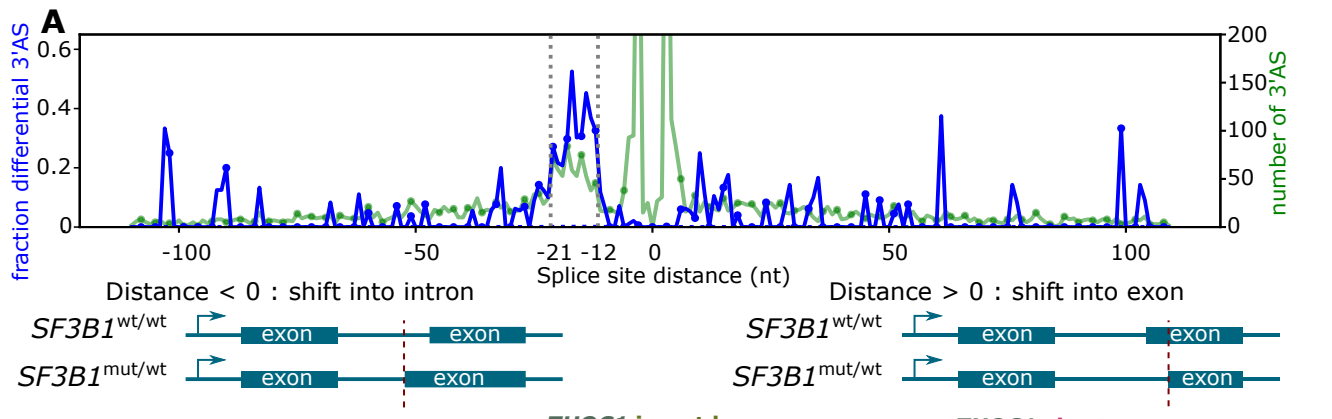
Zuallaert J, Godin F, Kim M, Soete A, Saeys Y, De Neve W. 2018. SpliceRover: interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics* **34**: 4180–4188.

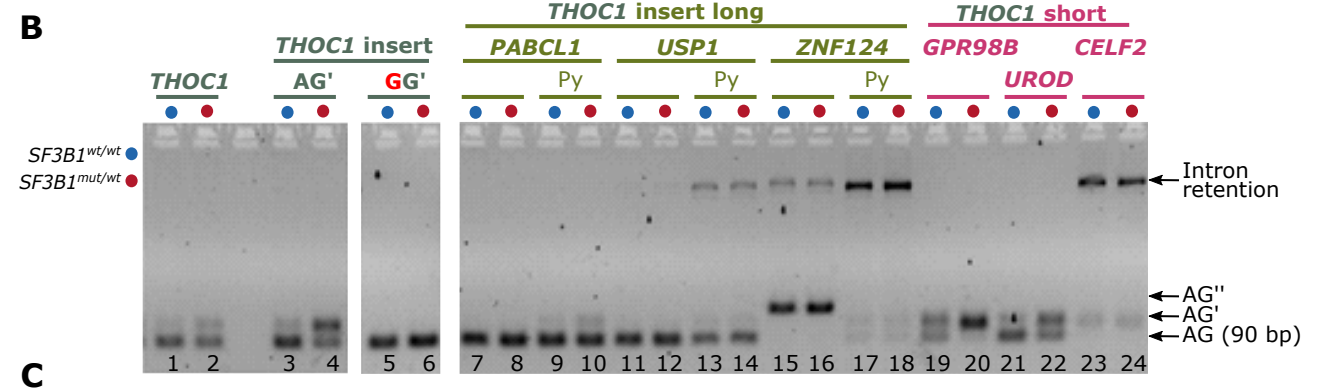

A
A284
Y623
R625
N626
H662
K666
Q699
K700
I704
G742
D781
D894
CLL
MDS
Cell lines
SF3B1-201 (FSM)
p14IR
H1 H2 H3 H4 H5 H6 H7 H8 H9 H10 H11 H12 H13 H14 H15 H16 H17 H18 H19 H20
0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300
B
Substantial transcripts expressed at ≥ 1 TPM
count
30,000
20,000
10,000
0
B cells
K562 SF3B1wt/wt
K562 SF3B1K700E/wt
Nalm6 SF3B1wt/wt
Nalm6 SF3B1H662Q/wt
CLL SF3B1wt/wt
CLL SF3B1mut/wt
MDS SF3B1wt/wt
MDS SF3B1mut/wt
Novelty class
Known (annotated):
Full splice match (FSM)
Novel (non-annotated):
Incomplete splice matches (ISM)
Novel in catalog (NIC)
Novel not in catalog (NNC)
Novel genes (NOVEL)
C
CLL SF3B1mut/wt
MDS SF3B1mut/wt
CLL SF3B1wt/wt
MDS SF3B1wt/wt
3,311
2,767
467
3,189
721
6,497
8,400
14,439
1,528
3,496
568
6,699
1,092
Nalm6 SF3B1H662Q/wt
Nalm6 SF3B1wt/wt
CLL SF3B1mut/wt
CLL SF3B1wt/wt
2,278
879
370
1,679
2,913
6,789
15,141
7,598
6,422
1,584
768
2,716
2,842
K562 SF3B1K700E/wt
MDS SF3B1mut/wt
K562 SF3B1wt/wt
MDS SF3B1wt/wt
1,583
5,811
920
3,969
1,820
7,343
12,661
7,377
1,882
2,563
444
5,975
891
D
Exons: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
SF3B1 (FSM)
SF3B1 (CLL)
SF3B1 (MDS)
Retained intron
Premature termination codon
p14 interacting region
HEAT
Anchor
900 nt / 300 aa
E
SF3B1 (FSM)
SF3B1 (CLL)
SF3B1 (MDS)
Percentage
75
50
25
0
CLL SF3B1wt/wt
CLL SF3B1mut/wt
MDS SF3B1wt/wt
MDS SF3B1mut/wt
Cell lines SF3B1wt/wt
Cell lines SF3B1mut/wt
B cells


A
ME
IR
ES
ALJ
AFJ
5'AS
3'AS
ASE type
0
5,000
10,000
15,000
Alternative splicing events (ASEs)
Novel: False True
B
Differential ASEs identified in each dataset (q-value < 0.1)
MDS 164
CLL 229
Cell lines
22
20
13
17
32
C
Cell lines vs CLL
R = 0.84
p = 1.1 × 10-46
CLL ΔPSI
Cell lines ΔPSI
Cell lines vs MDS
R = 0.84
p = 1.4 × 10-47
MDS ΔPSI
Cell lines ΔPSI
CLL vs MDS
R = 0.73
p = 7.9 × 10-41
MDS ΔPSI
CLL ΔPSI
D
Cell lines-specific
CLL-specific
MDS-specific
log(TPM + 0.1)
***
NS
NS.
Cell lines
CLL
MDS


A
down
up
ME
IR
ES
ALJ
AFJ
5'AS
3'AS
ASE type
6
7
189
24
30
43
43
48
27
15
9
8
71
255
200
100
0
100
200
300
Significant ASEs, q-value < 0.1
B
208 significant 3'AS events, q-value < 0.01
PSI
0
50
100
I
II
III
IV
CLL SF3B1wt/wt
CLL SF3B1wt/mut
MDS SF3B1wt/wt
MDS SF3B1wt/mut
Cell lines SF3B1wt/wt
Cell lines SF3B1wt/mut
C
CLL (3'AS)
MDS (3'AS)
Cell lines (3'AS)
PC2 (14.07 %)
PC1 (16.20 %)
PC2 (10.63 %)
PC1 (19.85 %)
PC2 (23.35 %)
PC1 (34.14 %)
B cells
D
SEPTIN6 3'AS
MAP3K7 3'AS
SETD4 3'AS
PRPF38A 3'AS
THOC1 3'AS
SF3B1mut/wt
SF3B1wt/wt
PSI
E
SEPTIN6
MAP3K7
SETD4
PRPF38A
THOC1
H20
100 bp
50 bp
K562-SF3B1wt/wt
K562-SF3B1K700E/wt
Nalm6-SF3B1wt/wt
Nalm6-SF3B1H662Q/wt
F
SF3B1wt
or
SF3B1K700E
+
Minigene
Exon 1
AG'
AG
Exon 2
HEK293FT
cDNA synthesis using minigene-specific oligos
PCR & gel electrophoresis
AG'
AG
G
SETD4
PRPF38A
THOC1
SEPTIN2
AG'
AG

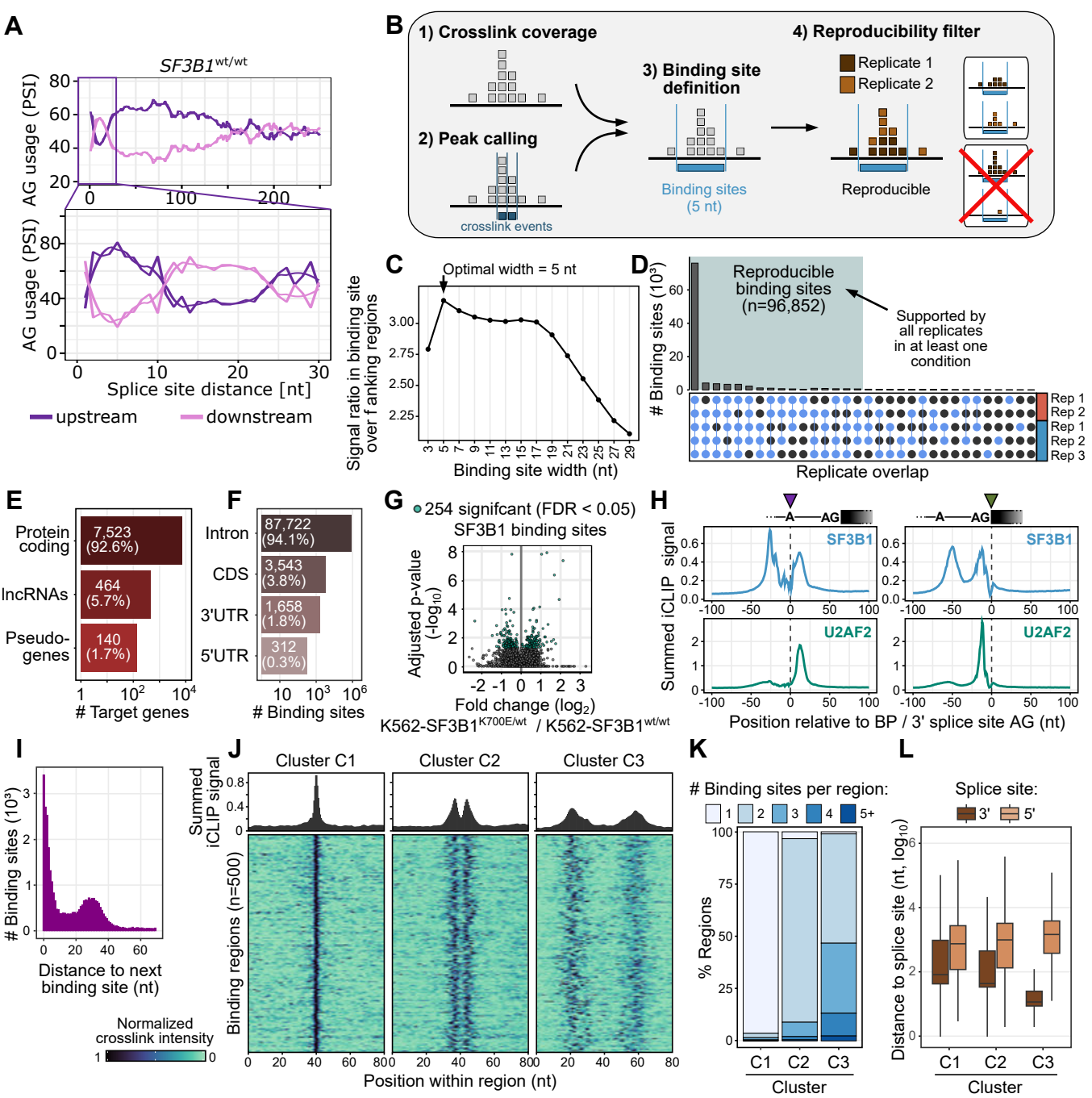

A
Known transcripts (FSM)
69.3%
3.6%
5.5%
0.4%
3.0%
14.0%
4.1%
Novel transcripts
29.6%
2.2%
11.0%
26.8%
20.1%
4.2%
6.1%
CDS match
Initiation match
De-novo
Non-coding
Stable
NMD
B
Differential events
120
80
40
0
ES
IR
ME
3'AS
5'AS
5'UTR
Disrupted start
In frame
Frameshift
Disrupted stop
3'UTR
Noncoding
Stable
NMD
C
SF3B1
Exons: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
SF3B1 (FSM)
SF3B1 (CLL)
retained intron
SF3B1 (MDS)
p14 interacting region
HEAT
Anchor
ORF start
ORF end
Premature stop codon
NMD
900 nt / 300 aa
D
PRPF38A
CLL SF3B1wt/wt
CLL SF3B1mut/wt
MDS SF3B1wt/wt
MDS SF3B1mut/wt
CL SF3B1wt/wt
CL SF3B1mut/wt
B cells
Relative expression [%]
0 25 50 75
FSM
3'AS
PRPF38A
Exons: 1 2 3 4 5 6 7 AG 8 9 10
300 nt / 100 aa
AG'
PRP38
PRP38_assoc
ORF start
ORF end
3'AS
E
MAP3K7
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
300 nt / 100 aa
Protein kinase
ORF start
ORF end
Translation re-initiation
Premature stop codon
uORF
3'AS
FSM 205
3'AS 205
FSM 202
3'AS 202
MAP3K7
CLL SF3B1wt/wt
CLL SF3B1mut/wt
MDS SF3B1wt/wt
MDS SF3B1mut/wt
CL SF3B1wt/wt
CL SF3B1mut/wt
B cells
0 20 40 60
Relative expression [%]

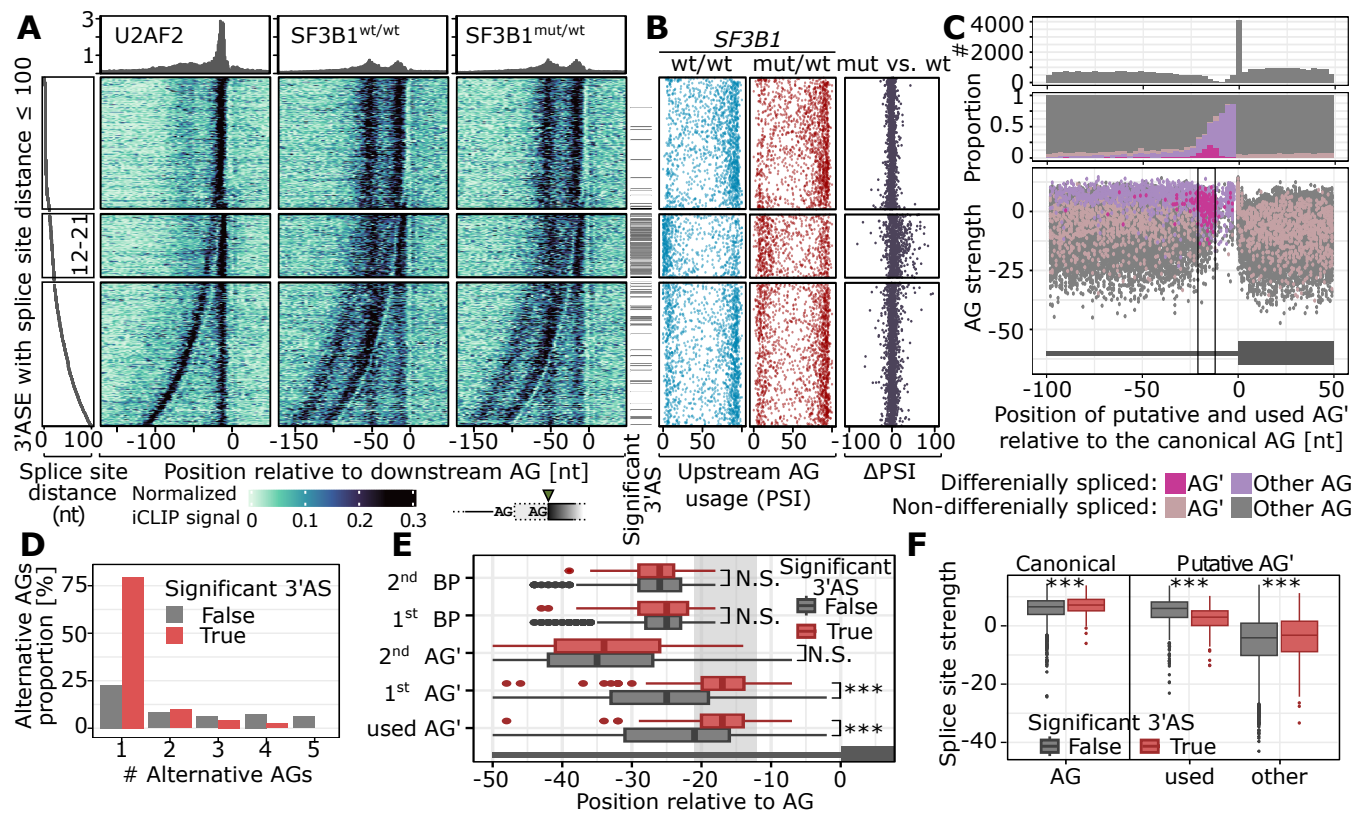




C

			THOC1: insertions between AG and AG'								
	THOC1	THOC1 AG'>GG	PABCL1	PABCL1 (Py)	USP1	USP1 (Py)	ZNF124	ZNF124 (Py)	GPR98B	UROD	CELF2
AG score	0.988	0.996	0.986	0.995	0.987	0.979	0.719	0.498	0.992	0.995	0.301
AG' score	0.704	/	0.848	0.507	0.210	0.660	0.952	0.928	0.665	0.857	0.902
AG–AG' distance	21	/	58	24	58	23	54	22	19	19	18


A
SF3B1wt/wt
AG usage (PSI)
Splice site distance [nt]
upstream
downstream
B
1) Crosslink coverage
2) Peak calling
crosslink events
3) Binding site definition
Binding sites (5 nt)
4) Reproducibility filter
Replicate 1
Replicate 2
Reproducible
C
Optimal width = 5 nt
Signal ratio in binding site over flanking regions
Binding site width (nt)
D
Reproducible binding sites (n=96,852)
Supported by all replicates in at least one condition
Binding sites (10³)
Rep 1
Rep 2
Rep 1
Rep 2
Rep 3
Replicate overlap
E
Protein coding
7,523 (92.6%)
lncRNAs
464 (5.7%)
Pseudo-genes
140 (1.7%)
Target genes
F
Intron
87,722 (94.1%)
CDS
3,543 (3.8%)
3'UTR
1,658 (1.8%)
5'UTR
312 (0.3%)
Binding sites
G
254 signifcant (FDR < 0.05) SF3B1 binding sites
Adjusted p-value (-log10)
Fold change (log2)
K562-SF3B1K700E/wt / K562-SF3B1wt/wt
H
SF3B1
U2AF2
Summed iCLIP signal
Position relative to BP / 3' splice site AG (nt)
I
Binding sites (10³)
Distance to next binding site (nt)
Normalized crosslink intensity
J
Cluster C1
Cluster C2
Cluster C3
Summed iCLIP signal
Binding regions (n=500)
Position within region (nt)
K
Binding sites per region:
1
2
3
4
5+
% Regions
Cluster
L
Splice site:
3'
5'
Distance to splice site (nt, log10)
Cluster


A
U2AF2
SF3B1wt/wt
SF3B1mut/wt
3'ASE with splice site distance ≤ 100
12-21
Splice site distance (nt)
Position relative to downstream AG [nt]
Normalized iCLIP signal
0 0.1 0.2 0.3
AG AG
B
SF3B1
wt/wt
mut/wt
mut vs. wt
Significant 3'AS
Upstream AG usage (PSI)
ΔPSI
C
4000 2000 0
Proportion 1 0.5 0
AG strength 0 -25 -50
-100 -50 0 50
Position of putative and used AG' relative to the canonical AG [nt]
Differenially spliced: AG' Other AG
Non-differenially spliced: AG' Other AG
D
Alternative AGs proportion [%]
75 50 25 0
Significant 3'AS
False
True
1 2 3 4 5
Alternative AGs
E
2nd BP
1st BP
2nd AG'
1st AG'
used AG'
N.S.
N.S.
N.S.
***
***
Significant 3'AS
False
True
-50 -40 -30 -20 -10 0
Position relative to AG
F
Canonical
Putative AG'
***
***
***
Splice site strength
0 -20 -40
Significant 3'AS
False
True
AG
used
other