HHS Public Access

Author manuscript

Oncogene. Author manuscript; available in PMC 2017 February 24.

Published in final edited form as:

Oncogene. 2017 February 23; 36(8): 1123–1133. doi:10.1038/onc.2016.279.

Cancer associated SF3B1 mutants recognize otherwise inaccessible cryptic 3' splice sites within RNA secondary structures

Anil K. Kesarwani[1,4], **Oscar Ramirez**[1,4], **Abhishek K. Gupta**[1], **Xiaodong Yang**[1], **Tushar Murthy**[2], **Alex C. Minella**[3], and **Manoj M. Pillai**[1,*]

[1]Section of Hematology, Yale Cancer Center and Yale University School of Medicine, New Haven, CT, USA

[2]Driskill Graduate Program, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

[3]Blood Research Institute, Blood Center of Wisconsin, Milwaukee WI, USA

Abstract

Recurrent mutations in core splicing factors have been reported in several clonal disorders, including cancers. Mutations in SF3B1, a component of the U2 splicing complex, are the most common. SF3B1 mutations are associated with aberrant pre-mRNA splicing using cryptic 3' splice sites (3'SS) but the mechanism of their selection is not clear. To understand how cryptic 3'SS are selected, we performed comprehensive analysis of transcriptome-wide changes to splicing and gene expression associated with SF3B1 mutations in patient samples as well as an experimental model of inducible expression. Hundreds of cryptic 3'SS were detectable across the genome in cells expressing mutant SF3B1. These 3'SS are typically sequestered within RNA secondary structures and poorly accessible compared to their corresponding canonical 3'SS. We hypothesized that these cryptic 3'SS are inaccessible during normal splicing catalysis and that this constraint is overcome in spliceosomes containing mutant SF3B1. This model of secondary structure-dependent selection of cryptic 3'SS was found across multiple clonal processes associated with SF3B1 mutations (myelodysplastic syndrome and chronic lymphocytic leukemia). We validated our model predictions in mini-gene splicing assays. Additionally, we found deregulated expression of proteins with relevant functions in splicing factor-related diseases both in association with aberrant splicing and without corresponding splicing changes. Our results show that SF3B1 mutations are associated with a distinct splicing program shared across multiple clonal processes and define a biochemical mechanism for altered 3'SS choice.

To whom correspondence should be addressed. Tel: 203-737-6403, Fax: 203-785-7232, Manoj.Pillai@Yale.edu.
[4]These authors contributed equally to the work.

CONFLICT OF INTERESTS

The authors have no conflicts of interests to declare.

AUTHOR CONTRIBUTIONS

AK analyzed next generation data and wrote the manuscript, OR performed experiments and wrote the manuscript, XY, AG and TM performed experiments, ACM designed some experiments, analyzed data and edited the manuscript, MMP initiated the study, performed experiments, analyzed data, provided supervision and wrote the manuscript with input from other authors.

Keywords

SF3B1; CANCER; MYELODYSPLASTIC SYNDROME; SPLICING

INTRODUCTION

Whole genome sequencing (WGS) has recently identified recurrent mutations in the core splicing machinery of several acquired clonal diseases including cancers[1]. Splicing mutations are most common in myelodysplastic syndromes (MDS), a group of clonal myeloid neoplasms[2,3] of the blood characterized by progressive bone marrow failure and genomic instability resulting in leukemic transformation[4–6]. Most of these mutations are restricted to four proteins (SF3B1, U2AF1, SRSF2 and ZRSR2), of which SF3B1 mutations are the most common. Splicing factor mutations have also been reported in other neoplasms including chronic lymphocytic leukemia (CLL)[7] and uveal melanoma[8], suggesting common underlying mechanisms of clonal evolution in splicing factor mutations. While it is logical to suspect the aberrant splicing of key transcripts are directly responsible for clonal evolution and hematopoietic dysplasia, the details of such a mechanism are not fully known.

SF3B1 is a component of the multi-protein complex SF3B and is thought to be important in stabilizing the base pairing of U2 snRNA to the branch point (BP)[9]. All disease-associated mutations of SF3B1 are within its 22 HEAT domains that form rod-like helical structures in its carboxyl-terminal region. SF3B1-mutant MDS is also associated with a specific phenotype called “ring sideroblasts”, which are maturing red cell precursors with a ring of iron-laden mitochondria around their nuclei[4]. Recently, multiple groups have described usage of novel or “cryptic” 3’ splice sites (3’SS) in patient samples with SF3B1 mutations[10,11]. These aberrantly spliced transcripts may have loss- or gain-of-function physiology, thus providing a link between SF3B1 mutations and disease biology. Two mechanisms have been recently proposed to understand this link. Based on computational modeling of the distance distribution of cryptic 3’SS in relation to their corresponding canonical 3’SS, DeBoever et al. proposed that mutant SF3B1 is able to recognize cryptic 3’SS located proximal to the sterically protected region downstream of the BP, without actually changing the BP choice itself[10]. Darman et al. proposed a different model by which the SF3B1-mutants have altered BP selectivity that lead to altered 3’ splice site choice[11]. In this manuscript, we provide evidence that these cryptic 3’SS are rendered inaccessible due to their presence within RNA secondary structures, while they are recognized by mutant SF3B1-containing spliceosomes. These computational models were verified in mini-gene splicing experiments. Our results provide a potential biochemical mechanism for altered 3’ splice site choice in SF3B1-mutant expressing cells and could ultimately provide insight into mechanisms of mutant SF3B1-associated MDS.

RESULTS

1. Aberrant splicing of transcripts using cryptic 3' splice site in cells expressing mutant SF3B1

Our analysis of mutant SF3B1-associated RNA splicing utilized two systems. Given that native human SF3B1 open reading frame (ORF) is toxic to E. coli[12], we synthesized a codon-optimized construct that allowed easily cloning (**Materials and Methods**). For transgenic expression of this construct, we developed an inducible lentiviral expression system with the hematopoietic cell line K562 and pInducer vector system by which we could limit the transgene expression levels close to that of the endogenous protein (Supplementary Figure 1). Cells induced to express SF3B1 wild-type or K700E are referred to as $SF3B1^{WT}$ and $SF3B1^{K700E}$ respectively. Paired-end RNA-Sequencing (RNA-Seq) was performed on total RNA depleted of ribosomal RNA (rRNA) using Illumina HiSeq2000 resulting in approximately 82×10^6 reads per library (detailed statistics shown in Supplementary Tables 1 & 2). rRNA depletion was employed instead of polyA selection to avoid loss of non-adenylated intermediaries. Mapping of paired-end RNA-Seq to the genome (hg19) resulted in a total of 182,190 splice junctions (SJ) from $SF3B1^{WT}$ and $SF3B1^{K700E}$ samples. When compared to 344,667 previously annotated SJ (Gencode v19), 75,857 were determined to be novel. SJ were then classified as those present only in $SF3B1^{WT}$, those present only in $SF3B1^{K700E}$, or shared between the two samples ($SF3B1^{WT}$ –$SF3B1^{K700E}$) (Supplementary Table 3) for downstream analysis. Similar analyses on patient samples (bone marrow derived CD34+ cells from 8 SF3B1-mutant MDS patients (variant allele frequency or VAF of 46–52% and 5 healthy controls) available from the GEO database (GSE63569)[13]. We found 430,052 SJ in this comparison of which 219,297 were novel.

We analyzed novel SJ to determine novel or "cryptic" 3'SS (defined as SJ associated with annotated 5'SS but novel 3'SS). 741 and 859 cryptic 3'SS were found in $SF3B1^{K700E}$ and MDS patient samples respectively (Supplementary File 1). These cryptic 3'SS were found to be widely distributed in their distance from corresponding canonical 3'SS (60% were upstream and 40% were downstream in $SF3B1^{K700E}$; 61% and 39% respectively in the MDS patient samples). Importantly, we noted an enrichment of cryptic 3'SS immediately upstream (10–35 nucleotide or nt) of the canonical 3'SS (Figure 1A), but no such pattern was evident in the downstream cryptic 3'SS (Supplementary Figure 2A). When the sequence motif in region upstream of cryptic 3'SS was compared to canonical 3'SS, the polypyrimidine tract (PPT) was noted to be interrupted with adenosines (Figure 1B), suggesting relatively weaker PPT. Full motif analysis of regions associated with upstream and downstream cryptic 3'SS is shown in Supplementary Figures 2B–D.

2. Cryptic 3'SS are less accessible compared to canonical counterparts

Although our results discussed above and previous reports[10,11,14] show that sequence upstream of cryptic and canonical 3'SS differ (PPTs enriched in adenosines for the former), this does not provide a direct mechanism for their selection. We hence considered other attributes of cryptic 3'SS including secondary structure. Secondary structure of nascent transcripts is a well-known determinant of splicing outcome[15]. Effect of secondary

structures have been demonstrated on the choice of 5'SS, 3'SS and BP across multiple species[16–18]. In eukaryotes, 3' SS selection is thought to involve 5' → 3' linear scanning of the sequence downstream from the BP, and the first AG downstream is selected[19]. However, AGs are skipped when they are too close to the BP (presumably since these are sterically protected) or if they are inaccessible within hairpin structures[20]. The enrichment of cryptic 3'SS immediately upstream of canonical 3'SS prompted us to check if they are hidden within secondary structures and hence not accessible normally during the 5' → 3' scanning in the second step of splicing. We first investigated the sequence around several candidate cryptic 3'SS for presence of secondary structures using Mfold[21]. As shown by representative examples in Figure 1C and Supplementary Figure 3, cryptic 3'SS are typically located within well-defined secondary structures that exclude their corresponding canonical 3'SS. This prompted us to test our hypothesis across the genome. Descriptive examples of cryptic 3'SS usage in SF3B1^{K700E} are shown in Figure 1D.

Since nascent pre-mRNA folds co-transcriptionally, multiple dynamically forming secondary structures are possible[22,23], by which the AG dinucleotide of a particular 3'SS can base-pair and hence be rendered inaccessible. One strategy to account for these scenarios is to include base-pairing probability (or corresponding accessibility of 3'SS) across multiple structures including optimal and sub-optimal[24,25]. Comparison of accessibility between cryptic and canonical 3'SS would be most meaningful in instances where they are both associated with a common BP. If our hypothesis were true, cryptic 3'SS would be less accessible than canonical during 5' → 3' scanning from the BP. However, BP in vertebrates are highly degenerate and hence cannot be computationally determined precisely for all 3'SS[26]. To computationally test our model, we hence identified those cryptic 3'SS which were reliably predicted to shared a BP with their corresponding canonical 3'SS using SVM_BP, a tool to predict BP in mammalian introns[27]. 275 such cryptic 3'SS in the patient dataset and 97 from SF3B1^{K700E} were identified (placed upstream of their canonical 3'SS). 109 and 59 such cryptic 3'SS with shared predicted BP were identified to be downstream of the canonical 3'SSin the patient sample and K562 datasets respectively. To compare relative accessibility, a sequence stretch downstream of the shared BP including the cryptic and canonical 3'SS and a further 15 nt downstream was selected (Figure 2A) and accessibility of cryptic and canonical 3'SS was calculated using RNAfold[28].

Accessibility of cryptic 3'SS was significantly reduced compared to their corresponding canonical counterparts (Figure 2B, Wilcoxon signed rank test; p = 2.20 E-16 for MDS patients and < 6.629E-08 for SF3B1^{K700E}). We performed similar analysis with 214 cryptic 3'SS from an independent dataset of CLL patients[10], which also showed a similar pattern of reduced accessibility (p =1.28 E-10). Importantly, similar results were obtained for cryptic 3'SS that were located downstream of canonical 3'SS with shared predicted BP (Supplementary Figure 4A). These analyses strongly support the hypothesis that cryptic 3'SS are protected within secondary structures in normal conditions, an effect that is seen regardless of whether cryptic 3'SS are located upstream or downstream of canonical counterparts, as long as they are in proximity.

Direct comparison of accessibility is not feasible for those cryptic/ canonical pairs that do not share a branch point (either because they are far from each other and hence unlikely to

share a BP, or no reliable BP predication could be made). Accessibilities calculated on independent sequence windows (starting from their respective BP to 15 nt downstream) are not expected to be different given that these splice sites are not in competition during the 5' → 3' scanning of the spliceosome. Consistent with this hypothesis, we found that their relative accessibilities were not different (Supplementary Figure 4B & C). Taken together, our results show that accessibility of cryptic 3'SS which share a BP with their canonical 3'SS is significantly lower than the canonical 3'SS. While the model can be extended to all cryptic 3'SS regardless of their distance from canonical 3'SS (detailed in Discussion section), rigorous computational validation is only possible in this subset. Given that core splice motifs critical to normal function are typically highly conserved, we also investigated if cryptic 3'SS are also conserved using PhyloP conservation score for 100 vertebrates[29] (Figure 2C). While canonical 3'SS showed high levels of conservation, cryptic 3'SS were poorly conserved across vertebrates suggesting that these sites do not have an evolutionarily preserved function.

To validate our computational model of 3'SS accessibility, we performed splicing assays using mini-gene constructs with different intronic inserts where splice forms from cryptic and canonical 3'SS could be differentiated by size resolution on agarose gel. (Figure 3A&B). A control intron (91 bp long, based on a mouse immunoglobulin heavy chain[30]) with canonical splice donor and acceptor sites was cloned between chicken Troponin T exonic elements (RG6 vector)[31] to create the control mini-gene (designated as I). Two additional constructs were designed: mini-gene II with a 12-bp hairpin upstream of the canonical 3'SS and mini-gene III which lacked the 5' side of the hairpin (no secondary structure is hence predicted to form). A cryptic AG was placed 14 bp upstream of mini-genes II and III. $SF3B1^{WT}$ or $SF3B1^{K700E}$ constructs were co-transfected with the mini-gene constructs into 293FT cells and RNA harvested after 48 hours (methodology detailed in Supplementary Methods). PCR based analysis of the cDNA as shown in Figure 3C revealed lower utilization of cryptic 3'SS within the hairpin (mini-gene II) in $SF3B1^{WT}$ expressing cells but higher utilization in $SF3B1^{K700E}$ cells. The upstream cryptic 3'SS in mini-gene III were preferentially used by both the $SF3B1^{WT}$ and $SF3B1^{K700E}$ expressing cells. Our results are thus in agreement with the hypothesis that upstream AGs typically not accessible within secondary structures are rendered accessible in $SF3B1^{K700E}$ expressing cells.

3. $SF3B1^{K700E}$ expression alters splicing pattern of hundreds of genes including those associated with cancer

Cryptic 3'SS as described above were determined by analysis of splice junctions, To determine alternative splicing events (skipped exon; SE, mutually exclusive exon; MXE, alternative 5' splice site; A5SS, alternative 3' splice site; A3SS and retained intron; RI), we implemented rMATS, an algorithm that takes into consideration exon coverage in addition to splice junction reads. From a comprehensive set of transcript annotations (assembled with Cufflinks[32]), we extracted and quantified the five common splicing events (Figure 4A and Supplementary Figure 5A&B). A total of 10,435 significant alternatively spliced events spread across 4,742 annotated genes were detected in $SF3B1^{WT}$ and $SF3B1^{K700E}$ samples, of which 3,886 events (in 2,356 genes) showed differential splicing (FDR ≤ 0.05, Supplementary File 2). Similar analysis in Healthy vs. MDS showed 75,903 events of which

4,788 events (in 2,823 genes) showed change in splicing (FDR ≤ 0.05) (Supplementary File 3). The most prevalent splicing event in both dataset comparisons is skipped exon (SE), encompassing about 56% ($SF3B1^{WT}$ vs. $SF3B1^{K700E}$) and 62% (Healthy vs. MDS) of their differentially spliced events (Figure 4B). The preponderance of SE as the predominant differentially spliced event (in excess of A3SS) is seemingly contrary to the primary role of SF3B1 in 3'SS selection. Both technical and biological reasons could explain this discrepancy: (1) Skipped exons are the most common alternative splicing event annotated in human transcriptome[33]. Given that rMATS analysis is biased towards existing annotations, this may result in a higher proportion of SE compared to other events. (2) Skipped exons may indirectly result from altered activity of other splicing regulators (such as PTB, HNRP and SR proteins), which are differentially spliced in our datasets (Supplementary File 2&3). Interestingly, a similar study of splicing changes induced by U2AF1 mutants also showed a preponderance of SE[34].

The ΔPSI (difference between "percentage spliced in" of $SF3B1^{WT}$ and $SF3B1^{K700E}$) for all significant events have a wide distribution (Figure 4C). When comparing K562-$SF3B1^{WT}$ and $SF3B1^{K700E}$, 39% of events had ΔPSI less than 10% (or |0.1|) while the corresponding proportion was 75% in patient samples suggesting that the demonstrable change in splicing isoform was modest for most transcripts. A list of genes with the 50 highest ΔPSI values is listed in Table 1. We performed unsupervised cluster analysis of inclusion level of the predominant splicing event, SE with at least ΔPSI of 10% (243 events in $SF3B1^{WT}$ vs. $SF3B1^{K700E}$ and 510 events in Healthy vs. MDS). In both comparisons, samples clustered along their SF3B1 mutational status confirming that the splicing events reproducibly linked to expression of mutant SF3B1 (Supplementary Figure 5C&D). Approximately 40% (953 of 2356) differentially spliced genes in $SF3B1^{WT}$ vs. $SF3B1^{K700E}$ were represented in the patient datasets, although only 6% of the specific events themselves were shared (Supplementary Figure 5E&F). The low concordance of specific splice events may be explained by the distinct transcriptomes of K562 cells and CD34-selected cells, as well as by differences between induced vs. stable expression systems. Nonetheless, those altered splicing events shared between the patient samples and our experimental model may be highly conserved across different cell types and, as such, may be important for pathogenesis of disease related to recurrent HEAT domain mutations in SF3B1. Several genes with relevance to cancer and MDS biology were noted to undergo alternative splicing (Supplementary File 2 and 3). However, most including *EZH2, NRAS* and *GATA1* contained events with low ΔPSI values. We experimentally validated differentially spliced events in four genes relevant to cancer and MDS biology (*BCL2L12*[35], *DROSHA*[36], *BOLA3*[37] and *IMMT*[38]) that were present in both the $SF3B1^{WT}$ vs. $SF3B1^{K700E}$ and MDS patients vs. healthy comparisons, by quantitative PCR (Figure 4D).

4. $SF3B1^{K700E}$ expression leads to differential expression of genes both directly and indirectly

To investigate how these differentially spliced genes may contribute to MDS pathophysiology, we performed pathway analysis (Ingenuity Pathway Analysis or IPA) and determined the biological pathways enriched in these datasets (Figure 5A and Supplementary File 4). Both the datasets showed significant enrichment for pathways

directing functions including protein quality control, mitochondrial function, cell death and proliferation and regulation of gene expression. One of these genes, UBE2I encodes the SUMO E2 ligase Ubc9[39] and was found to be differentially spliced (skipped exon in the 5' UTR that results in isoforms that differ in 5'UTR) (Supplementary Files 2 & 3 and Figure 5B). Although the total transcript level was not different (Figure 5C), protein levels in $SF3B1^{K700E}$ cells were found to be lower than $SF3B1^{WT}$ (Figure 5D), suggesting that the mechanism of differential gene expression may be post-transcriptional (regulatory elements in 5'UTR have been reported to regulate rate of translation[40]).

SLC25A37 or Mitoferrin 1 is a critical mitochondrial iron transporter, the loss of which results in severe anemia in experimental models[41]. Mitoferrin 1 has been reported to be both alternatively spliced and differentially expressed in SF3B1-mutant MDS[13,42]. We noted Mitoferrin 1 demonstrated differential splicing in $SF3B1^{K700E}$ as well as in MDS patient samples, albeit with only small differences in splicing ratio (Figure 5E&F). Interestingly, $SF3B1^{K700E}$ cells had higher total transcript levels but reduced protein levels of Mitoferrin 1 when compared to $SF3B1^{WT}$cells (Figure 5G–I, Western blotting for protein level was performed in MEL cells due to non-availability of reliable antibody against the human epitope). Mitoferrin 1 is known to be regulated post-translationally[43], hence the higher transcript levels may represent a compensatory increase to its degradation.

A small set of genes was found differentially expressed in $SF3B1^{K700E}$ cells compared to $SF3B1^{WT}$ cells. (Table 2 and Supplementary File 5). This list included several genes previously known to be important in MDS and cancer biology, including EGR1, (an immediate early transcription factor, the loss of which can lead to HSC expansion[44]) and DLK1 (a member of the EGF-like family of proteins with both transmembrane and secreted forms[45] known to be down-regulated in the RARS subtype of MDS[46]). Differential transcript and protein levels for EGR1 and DLK1 are shown in Supplementary Figure 6. Given that they are not alternatively spliced, it is likely that their differential expression is indirectly related to mutant SF3B1 expression. To discount mutant specific and cell line specific effects, our results were also confirmed for other HEAT domain mutants reported in MDS ($SF3B1^{K666N}$ and $SF3B1^{K666R}$) and in the murine MEL erythroleukemia cell line (Supplementary Figures 7 & 8).

Elimination of aberrantly spliced isoforms by nonsense-mediated decay (NMD) surveillance has been invoked as a mechanism by which splicing mutations may lead to changes in protein expression[47]. Potential for transcript isoforms to trigger NMD can be computationally predicted (premature termination codon or PTC, long 3' UTR, splice junction > 50 nt downstream of the stop codon). However, for NMD to be invoked as a potential mechanism for differential gene expression, a change in transcript abundance at steady state is expected. Among the few differentially expressed genes between $SF3B1^{WT}$ and $SF3B1^{K700E}$ cells (Table 2), only two (Mitoferrin 1 and HDAC7) show known NMD triggering features described above (Supplementary File 2&3). We determined if inhibition of NMD (by treatment with cycloheximide[48]) would increase the relative abundance of Mitoferrin 1 isoforms predicted to be NMD targets. We did not observe a significant change in these isoform levels assayed by quantitative RT-PCR (Supplementary Figure 9),

suggesting that mechanisms other than NMD may play a role in differential expression of genes including Mitoferrin 1 in SF3B1^{K700E} cells.

DISCUSSION

In this manuscript, we propose a new model for splicing changes associated with recurrent mutations in the HEAT domains of SF3B1 described in multiple clonal processes including MDS, CLL, uveal melanoma and breast cancer. Previous studies have described changes in 3'SS selection in cells that express mutant SF3B1[10,11]. Our results are in general agreement with these reports with regards to the distribution of cryptic 3'SS and associated weak PPT. In addition, our analysis shows that cryptic 3'SS used by SF3B1^{K700E} are typically sequestered within secondary structures and not accessible to nucleophilic attack in wild-type cells. SF3B1^{K700E} expression overcomes this constraint rendering them accessible (Figure 6).

Three recent studies have proposed divergent mechanistic models to explain how mutant SF3B1 expression leads to cryptic 3'SS usage. The first by DeBoever et al. proposed that mutant SF3B1 is able to recognize cryptic 3'SS by overcoming the usual steric hindrance in a region immediately downstream of the branch point (BP)[10]. Specifically, they discounted altered BP selection as the cause for cryptic 3'SS selection. Two limitations for the model were: (1) it could not explain cryptic 3'SS that fell farther away from this sterically protected region; and (2) the model was not supported with direct experimental validation. A second model based on multiple SF3B1-mutant cancer samples[11] (and a more recent similar study in SF3B1-mutant melanoma[14]) proposed that U2 spliceosomes with mutant SF3B1 alters the BP choice, which then results in altered 3'SS selection. Validation of this model was performed by mutational analysis of putative BP in a mini-gene constructed from candidate intronic segments. Some limitations of this model include: (1) more widespread 3'SS usage would be expected with altered BP choice (without the enrichment upstream of canonical sites); and (2) recent genome-wide mapping of human BP have shown that a third of exons have multiple potential BP, most of which are within a few base pairs of each other[26]. This strongly suggests that BP usage is highly redundant during splicing and is neither necessary nor sufficient for alternative 3'SS usage.

The model we propose is based on the well-described relationship of secondary structure constraints to splice site choice. Accordingly, cryptic 3'SS are not used normally as the spliceosome scans in the 5' → 3' direction since they are base-paired within secondary structures rendering them inaccessible. This difference in accessibility holds true regardless of the cryptic 3'SS being placed upstream or downstream of the canonical 3'SS. Our model also accommodates cryptic 3'SS that are distant (> 50 nt) from their canonical sites and hence likely to be associated with a different BP. Accordingly, these distal cryptic 3'SS are likely independent units (with distinct 3'SS, PPT and BP), but not used normally due to poor accessibility. They are however not in competition with their corresponding canonical 3'SS. In mutant SF3B1-containing spliceosomes, these cryptic 3'SS are recognized and used. Although U2AF assembles at 3'SS early during the splicing process, final selection of the acceptor site is now thought to occur later by a 5→3' scanning process downstream of the BP[19,49]. Importantly, our model does not rule out a change in BP choice for cryptic 3'SS

selection, rather it suggests that intrinsic features of the cryptic 3'SS (accessibility) determines their selection. Although computational and mini-gene splicing assays support our model, further validation will come from comprehensive charting of genome-wide BP choice in wild-type and mutant-SF3B1 expressing cells (through direct sequencing of lariat junctions). Cryptic 3'SS were also not found to be conserved across vertebrate species suggesting that they are unlikely to have a role in normal physiology. This may also explain recent reports that showed a lack of significant phenotype analogous to the human disease in mice expressing mutant SF3B1, despite having a similar pattern of cryptic 3'SS across the genome,[50,51] as transcripts with cryptic 3'SS in humans are unlikely to have the a homologous cryptic 3'SS in mice.

How mutant SF3B1-containing U2 spliceosomes select cryptic 3'SS within secondary structures is not clear. Proteins interacting with the C-terminal HEAT domains of SF3B1 are not known, but it is reasonable to speculate that the mutation could change its interaction with one or more of factors involved in 3'SS selection. This may include RNA helicases, which are essential for unwinding of both snRNA and target pre-mRNA during splicing[52]. Interestingly, recent descriptions of mutations in DDX41 (a dead box RNA helicase) and splicing anomalies in patients affected by these somatic mutations underscore a direct connection between pre-mRNA secondary structure and splicing in the context of MDS[53]. A separate potential mechanism is based on recently described interaction between SF3B1 and chromatin[54]. Given that splicing is co-transcriptional, rate of transcription dictates how nascent RNA folds into secondary structures. It is hence possible that mutant SF3B1 changes transcription rate such that the second step of splicing is completed before cryptic 3'SS are sequestered within dynamic secondary structure. Precise details of 3'SS choice in mutant SF3B1 cells will likely emerge out of studies that define the direct targets of SF3B1 binding (such as CLIP-Seq), crystal structure of SF3B1 and complete charting of BPs (lariat sequencing).

Differentially spliced genes are enriched for biological pathways of relevance to MDS and cancer biology. Interestingly, differential expression at the protein level could be demonstrable for alternatively spliced genes such as UBE2I although their transcript levels were not different. Conversely, some genes (such as EGR1 and DLK1) were not aberrantly spliced although differentially expressed at the transcript and protein levels. NMD of aberrantly spliced transcripts has been invoked as a potential mechanism that links splicing factor mutations and disease[11,55]; in our analysis NMD was not clearly demonstrable in differentially expressed genes including Mitoferrin 1. This suggests that the precise link between aberrant splicing and disease physiology is complex and may not be readily evident from transcriptome analysis for splicing and gene expression.

In summary, our results provide a model for understanding how mutant SF3B1 expression may lead to cryptic 3'SS selection. It forms the basis for further biochemical work that will define how the mutant protein accomplishes recognition of the inaccessible 3'SS and identify detailed mechanisms of tumorigenesis associated with mutant splicing factor expression.

MATERIALS AND METHODS

1. Cloning of SF3B1, cell culture, vector construction and viral vector production

Native SF3B1 sequences are toxic to *E. coli* as reported previously[12]. Sequence of human SF3B1 open reading frame (ORF) was optimized by in-house software provided by Genscript, synthesized and cloned into pUC57, and subsequently cloned to other expression vectors with an N-terminal FLAG-tag. Mutant SF3B1 constructs (for expression of K700E, K666N and K666R) were generated by site-directed mutagenesis as described in Supplementary Methods. Unless otherwise specified, mutant SF3B1 denotes $SF3B1^{K700E}$ (lysine to glutamic acid substitution at position 700, the most common mutation in MDS) while $SF3B1^{WT}$ denotes the unmutated wild-type construct. K562 and MEL were grown in RPMI-1640 supplemented with 10% fetal calf serum. 293FT cells were grown in DMEM supplemented with 10% cosmic calf serum (Hyclone). K562 cells were used as the experimental model given human CD34+ cell transductions resulted in low (~ 5%) efficiencies and an inability to expand in culture (likely due to toxicity from enforced expression). A single cistron doxycycline inducible viral vector model (pInducer 20 or 21 vectors, Supplementary Figure 1A &B), which allowed for titration of induced transgene levels by varying concentrations of doxycycline. Inducible lentiviral constructs were made by first cloning the SF3B1 constructs in to pENTR4 (Addgene plasmid 17424) vector and transferring to destination vectors (pInducer20 and pInducer 21 (Addgene plasmids 44012 and 46948 respectively[56]). Lentiviral vectors were prepared by calcium phosphate transfection of 293FT cells with viral plasmid constructs along with helper plasmids pdelta8.92 and pCMV-VSVG as previously described[57]. Viral supernatants were concentrated using PEG6000. Cells were transduced with viral concentrates and selected by either flow sorting for GFP expression (pinducer21 transduced cells) or Geneticin (500 ug/ml for 10–14 days). Inducible expression of FLAG-SF3B1 and total SF3B1 were separately determined by Western blot and levels compared to ensure that inducible transgene expression remained in the physiological range (Supplementary Figure 1C&D), Two single cell clones were identified for each transgene with comparable levels of inducible expression and subsequently studied. Primary analysis described in this manuscript are for $SF3B1^{WT}$ and SF3B1-K700E induced with doxycycline (referred henceforth as $SF3B1^{WT}$ and ^{K700E} respectively). Induction was performed for 48 hours with 1 ug/ml of doxycycline. For NMD analysis, cells were treated with cycloheximide (100 ug/ml) for 30 minutes or DMSO (control).

2. RNA sequencing and analysis

Total RNA was extracted from approximately 5 million cultured cells using RNEasy mini-kit. Ribosomal RNA (rRNA) depletion was performed by RNAseH treatment as previously described[58]. Strand-specific (dUTP protocol) cDNA libraries with barcodes suitable for multiplexed Illumina sequencing were prepared in duplicates for each condition (wild-type uninduced ($SF3B1^{WT-}$), wild-type induced ($SF3B1^{WT}$), K700E or mutant uninduced ($SF3B1^{K700E-}$) and mutant induced ($SF3B1^{K700E}$) induced. Pooled libraries in equimolar ratios (4 libraries per lane) were sequenced on the Illumina HiSeq2000, 2 × 76 (paired-end) per sample. Paired-end sequencing data of 8 MDS patients with SF3B1 mutations and 5 healthy controls were downloaded from GEO database (GSE63569)[13]. Details of informatic

analyses, including packages used, custom scripts generated are in Supplementary Methods. Custom scripts written for the analysis have been deposited at https://github.com/pillailab under the repository “SF3B1_Splicing”.

3. Reverse Transcription, PCR and mini-gene splicing assay

cDNA was prepared from total RNA depleted of genomic DNA (on-column DNAse treatment) and SuperscriptIII reverse transcriptase (Life Technologies) using random hexamers. Real-Time quantitative PCR was performed using paired oligos and Sybr Fast (KAPA Biosystems) 2× mastermix on a Bio-rad CFX96 real time PCR cycler. To determine splice isoforms, cDNA was amplified with suitable primers and resolved on 3% Metaphor agarose. Mini-gene splicing assays were performed using constructs cloned in to the RG6 vector[31] (details in Supplementary Methods).

4. Western Blot and ELISA

Protein detection and quantification were accomplished by western analysis using standard reagents and methodology as previously reported[57]. Specific conditions for blocking, primary and secondary antibody binding are described in Supplementary Methods. Immunoreactive bands were visualized using Pierce enhanced chemiluminescence (ECL) substrate (Life Technologies). ELISA for soluble DLK1 was performed using DLK1 duo – set ELISA kit (RnD Biosystems) per manufacturer’s instruction and read on a Biotek Synergy 2 plate reader.

5. Data access

Sequencing files generated from this work have been deposited to the Gene Expression Omnibus (GEO) under GSE70959.

6. Statistical Considerations

RNA-Seq was performed in duplicates as per ENCODE guidelines (https://genome.ucsc.edu/ENCODE/experiment_guidelines.html). Detailed statistical analytical parameters for RNA-Seq and splicing are detailed in Supplementary Methods. Individual experiments including PCR, Western blots and ELISA were performed in biological replicates of three. The results of each experiment are reported as the mean of experimental replicates. Error bars represent the s.d. from the mean. If not otherwise indicated, pairwise comparisons were analyzed using the unpaired two-sided *t*-test (GraphPad Prism). P-value less than 0.05 was considered significant.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Work was funded in part by the National Institutes of Health (R01 HL104070 to MMP, and R01 HL098608 to ACM). We thank Barry Paw (Boston Children’s Hospital, Boston MA) for anti-Mitoferrin antibody, Thomas A Cooper (Baylor College of Medicine, Houston TX) for the RG6 plasmid and Karla Neugebauer (Yale University) for helpful suggestions. We also acknowledge The Yale Center for Genome Analysis (YCGA) for high throughput

sequencing and the Yale University Biomedical High Performance Computing Center for use of compute clusters to run bioinformatics analysis.

REFERENCES

1. Yoshida K, Ogawa S. Splicing factor mutations and cancer. Wiley Interdiscip Rev RNA. 2014; 5:445–459. [PubMed: 24523246]
2. Greenberg PL, Young NS, Gattermann N. Myelodysplastic syndromes. Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program. 2002:136–161.
3. Bejar R, Steensma DP. Recent developments in myelodysplastic syndromes. Blood. 2014; 124:2793–2803. [PubMed: 25237199]
4. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. N Engl J Med. 2011; 365:1384–1395. [PubMed: 21995386]
5. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011; 478:64–69. [PubMed: 21909114]
6. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. Nat Genet. 2012; 44:53–57.
7. Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med. 2011; 365:2497–2506. [PubMed: 22150006]
8. Harbour JW, Roberson EDO, Anbunathan H, Onken MD, Worley LA, Bowcock AM. Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. Nat Genet. 2013; 45:133–135. [PubMed: 23313955]
9. Gozani O, Feld R, Reed R. 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. 1996; 10:233–243. [PubMed: 8566756]
10. DeBoever C, Ghia EM, Shepard PJ, Rassenti L, Barrett CL, Jepsen K, et al. Transcriptome sequencing reveals potential mechanism of cryptic 3' splice site selection in SF3B1-mutated cancers. PLoS Comput Biol. 2015; 11:e1004105. [PubMed: 25768983]
11. Darman RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, et al. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3? Splice Site Selection through Use of a Different Branch Point. Cell Rep. 2015; 13:1033–1045. [PubMed: 26565915]
12. Wang C, Chua K, Seghezzi W, Lees E, Gozani O, Reed R. Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis. Genes Dev. 1998; 12:1409–1414. [PubMed: 9585501]
13. Dolatshad H, Pellagatti A, Fernandez-Mercado M, Yip BH, Malcovati L, Attwood M, et al. Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells. Leukemia. 2014
14. Alsafadi S, Houy A, Battistella A, Popova T, Wassef M, Henry E, et al. Cancer-associated SF3B1 mutations affect alternative splicing by promoting alternative branchpoint usage. Nat Commun. 2016; 7:10615. [PubMed: 26842708]
15. Buratti E, Baralle FE. Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. Mol Cell Biol. 2004; 24:10505–10514. [PubMed: 15572659]
16. Meyer M, Plass M, Pérez-Valle J, Eyras E, Vilardell J. Deciphering 3′ss Selection in the Yeast Genome Reveals an RNA Thermosensor that Mediates Alternative Splicing. Molecular Cell. 2011; 43:1033–1039. [PubMed: 21925391]
17. Vořechovský I. Aberrant 3′ splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. Nucl Acids Res. 2006; 34:4630–4641. [PubMed: 16963498]
18. Warf MB, Berglund JA. Role of RNA structure in regulating pre-mRNA splicing. Trends in Biochemical Sciences. 2010; 35:169–178. [PubMed: 19959365]
19. Chua K, Reed R. An Upstream AG Determines Whether a Downstream AG Is Selected during Catalytic Step II of Splicing. Mol Cell Biol. 2001; 21:1509–1514. [PubMed: 11238888]

20. Smith CW, Chu TT, Nadal-Ginard B. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. Mol Cell Biol. 1993; 13:4939–4952. [PubMed: 8336728]
21. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003; 31:3406–3415. [PubMed: 12824337]
22. Mahen EM, Watson PY, Cottrell JW, Fedor MJ. mRNA secondary structures fold sequentially but exchange rapidly in vivo. PLoS Biol. 2010; 8:e1000307. [PubMed: 20161716]
23. Bevilacqua PC, Blose JM. Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. Annu Rev Phys Chem. 2008; 59:79–103. [PubMed: 17937599]
24. Plass M, Codony-Servat C, Ferreira PG, Vilardell J, Eyras E. RNA secondary structure mediates alternative 3′ss selection in Saccharomyces cerevisiae. RNA. 2012; 18:1103–1115. [PubMed: 22539526]
25. Plass M, Eyras E. Approaches to link RNA secondary structures with splicing regulation. Methods Mol Biol. 2014; 1126:341–356. [PubMed: 24549676]
26. Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, et al. Genome-wide discovery of human splicing branchpoints. Genome Res. 2015; 25:290–303. [PubMed: 25561518]
27. Corvelo A, Hallegger M, Smith CWJ, Eyras E. Genome-wide association between branch point properties and alternative splicing. PLoS Comput Biol. 2010; 6:e1001016. [PubMed: 21124863]
28. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL. The Vienna RNA Websuite. Nucleic Acids Res. 2008; 36:W70–W74. [PubMed: 18424795]
29. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20:110–121. [PubMed: 19858363]
30. Lai Y, Yue Y, Liu M, Duan D. Synthetic Intron Improves Transduction Efficiency of Trans-Splicing Adeno-Associated Viral Vectors. Human Gene Therapy. 2006; 17:1036–1042. [PubMed: 17007565]
31. Orengo JP, Bundman D, Cooper TA. A bichromatic fluorescent reporter for cell-based screens of alternative splicing. Nucleic Acids Res. 2006; 34:e148. [PubMed: 17142220]
32. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012; 7:562–578. [PubMed: 22383036]
33. Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, et al. Comparative analysis of the transcriptome across distant species. Nature. 2014; 512:445–448. [PubMed: 25164755]
34. Ilagan JO, Ramakrishnan A, Hayes B, Murphy ME, Zebari AS, Bradley P, et al. U2AF1 mutations alter splice site recognition in hematological malignancies. Genome Res. 2015; 25:14–26. [PubMed: 25267526]
35. Kontos CK, Scorilas A. Molecular cloning of novel alternatively spliced variants of BCL2L12, a new member of the BCL2 gene family, and their expression analysis in cancer cells. Gene. 2012; 505:153–166. [PubMed: 22664385]
36. Merritt WM, Lin YG, Han LY, Kamat AA, Spannuth WA, Schmandt R, et al. Dicer, Drosha, and outcomes in patients with ovarian cancer. N Engl J Med. 2008; 359:2641–2650. [PubMed: 19092150]
37. Cameron JM, Janer A, Levandovskiy V, Mackay N, Rouault TA, Tong W-H, et al. Mutations in iron-sulfur cluster scaffold genes NFU1 and BOLA3 cause a fatal deficiency of multiple respiratory chain and 2-oxoacid dehydrogenase enzymes. Am J Hum Genet. 2011; 89:486–495. [PubMed: 21944046]
38. Okeyo-Owuor T, White BS, Chatrikhi R, Mohan DR, Kim S, Griffith M, et al. U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. Leukemia. 2015; 29:909–917. [PubMed: 25311244]
39. Kaganovich D, Kopito R, Frydman J. Misfolded proteins partition between two distinct quality control compartments. Nature. 2008; 454:1088–1095. [PubMed: 18756251]
40. Floor SN, Doudna JA. Tunable protein synthesis by transcript isoforms in human cells. Elife. 2016; 5
41. Shaw GC, Cope JJ, Li L, Corson K, Hersey C, Ackermann GE, et al. Mitoferrin is essential for erythroid iron assimilation. Nature. 2006; 440:96–100. [PubMed: 16511496]

42. Visconte V, Rogers HJ, Singh J, Barnard J, Bupathi M, Traina F, et al. SF3B1 haploinsufficiency leads to formation of ring sideroblasts in myelodysplastic syndromes. Blood. 2012; 120:3173–3186. [PubMed: 22826563]
43. Paradkar PN, Zumbrennen KB, Paw BH, Ward DM, Kaplan J. Regulation of Mitochondrial Iron Import through Differential Turnover of Mitoferrin 1 and Mitoferrin 2. Mol Cell Biol. 2009; 29:1007–1016. [PubMed: 19075006]
44. Min IM, Pietramaggiori G, Kim FS, Passegué E, Stevenson KE, Wagers AJ. The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. Cell Stem Cell. 2008; 2:380–391. [PubMed: 18397757]
45. Moon YS, Smas CM, Lee K, Villena JA, Kim K-H, Yun EJ, et al. Mice lacking paternally expressed Pref-1/Dlk1 display growth retardation and accelerated adiposity. Mol Cell Biol. 2002; 22:5585–5592. [PubMed: 12101250]
46. Sakajiri S, O'kelly J, Yin D, Miller CW, Hofmann WK, Oshimi K, et al. Dlk1 in normal and abnormal hematopoiesis. Leukemia. 2005; 19:1404–1410. [PubMed: 15959531]
47. Popp MW-L, Maquat LE. The dharma of nonsense-mediated mRNA decay in Mammalian cells. Mol Cells. 2014; 37:1–8. [PubMed: 24552703]
48. Ishigaki Y, Li X, Serin G, Maquat LE. Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. Cell. 2001; 106:607–617. [PubMed: 11551508]
49. Horowitz DS. The mechanism of the second step of pre-mRNA splicing. Wiley Interdiscip Rev RNA. 2012; 3:331–350. [PubMed: 22012849]
50. Obeng EA, McConkey ME, Campagna D, Schneider RK, Chen MC, Schmidt PJ, et al. Mutant Splicing Factor 3b Subunit 1 (SF3B1) Causes Dysregulated Erythropoiesis and a Stem Cell Disadvantage. Blood. 2014; 124:828–828. [PubMed: 25104859]
51. Mupo A, Sathiaseelan V, Seiler M, Kent D, Peng S, Bautista R, et al. Sf3b1 K700E Mutation Impairs Pre-mRNA Splicing and Definitive Hematopoiesis in a Conditional Knock-in Mouse Model. Blood. 2015; 126:140–140.
52. Cordin O, Beggs JD. RNA helicases in splicing. RNA Biol. 2013; 10:83–95. [PubMed: 23229095]
53. Polprasert C, Schulze I, Sekeres MA, Makishima H, Przychodzen B, Hosono N, et al. Inherited and Somatic Defects in DDX41 in Myeloid Neoplasms. Cancer Cell. 2015; 27:658–670. [PubMed: 25920683]
54. Kfir N, Lev-Maor G, Glaich O, Alajem A, Datta A, Sze SK, et al. SF3B1 Association with Chromatin Determines Splicing Outcomes. Cell Rep. 2015
55. Kim E, Ilagan JO, Liang Y, Daubner GM, Lee SC-W, Ramakrishnan A, et al. SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. Cancer Cell. 2015; 27:617–630. [PubMed: 25965569]
56. Meerbrey KL, Hu G, Kessler JD, Roarty K, Li MZ, Fang JE, et al. The pINDUCER lentiviral toolkit for inducible RNA interference in vitro and in vivo. PNAS. 2011; 108:3665–3670. [PubMed: 21307310]
57. Balakrishnan I, Yang X, Brown J, Ramakrishnan A, Torok-Storb B, Kabos P, et al. Genome-wide analysis of miRNA-mRNA interactions in marrow stromal cells. Stem Cells. 2013
58. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods. 2013; 10:623–629. [PubMed: 23685885]

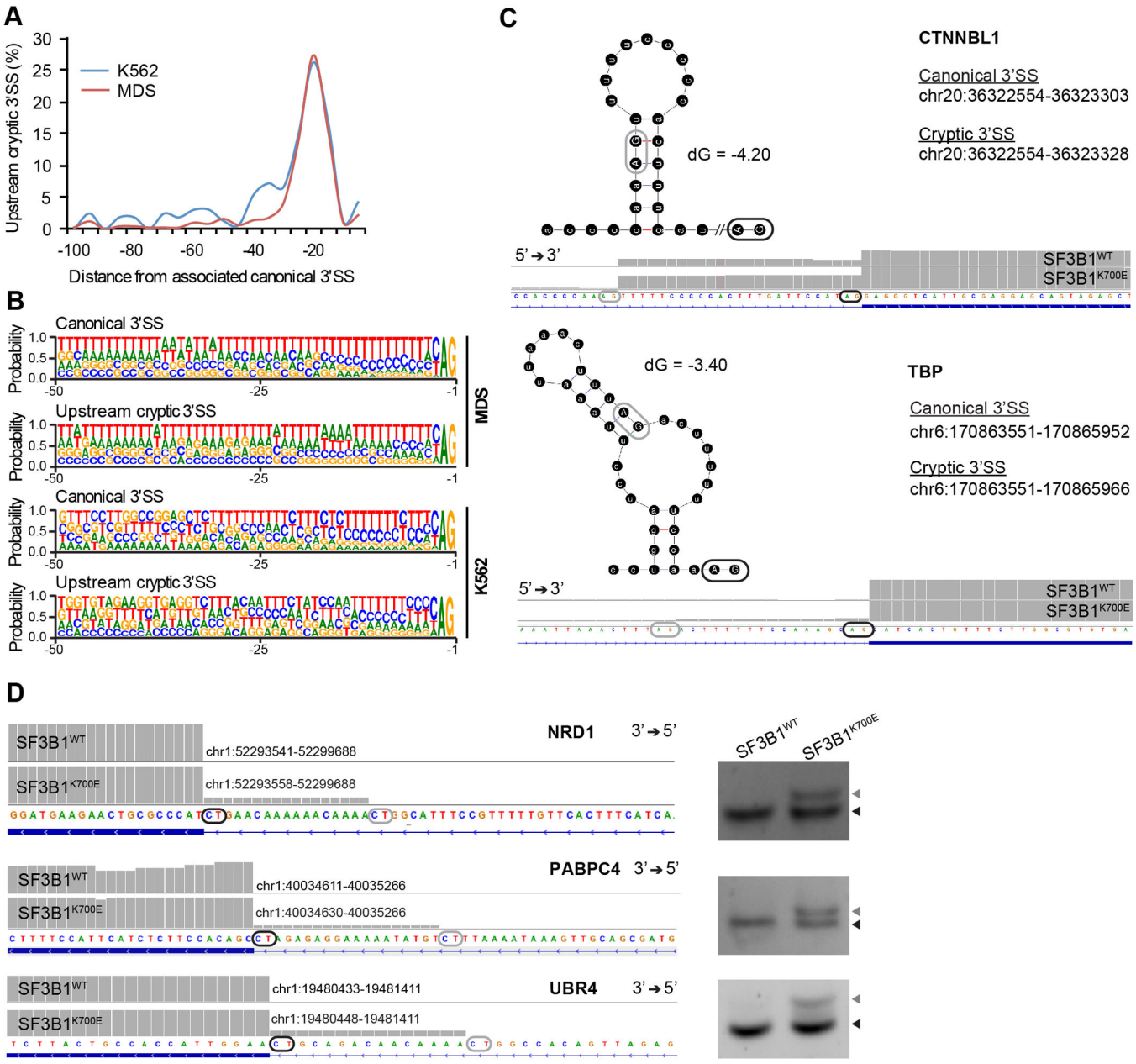

Figure 1. Features of cryptic 3'SS

(**A**) Distribution of cryptic 3'SS located upstream of their canonical 3'SS in MDS and SF3B1^{K700E} expressing K562 cells. Distance from canonical 3'SS is shown in 5 nucleotide bins. (**B**) Sequence motif in the region 50 nucleotides upstream of canonical 3'SS or cryptic 3'SS in MDS and K562 samples. (**C**) Representative examples of cryptic 3'SS inaccessible in secondary structure. Shown are the gene names and genomic coordinates of canonical and cryptic 3'SS, RNA-seq read support (IGV plot adjusted to same scale for SF3B1^{WT} and SF3B1^{K700E}) and secondary structure (Mfold plot). The dinucleotides surrounded by black and grey rounded-rectangles denote canonical and cryptic 3'SS respectively in the IGV plots and secondary structures. The orientation of the gene is shown, and the free energy (delta G) of structure as 'dG'. For better accommodation of structures in this figure, few bases near

'//' sign were removed after plotted by Mfold. (**D**) Three examples of cryptic 3'SS in $SF3B1^{K700E}$ cells co-amplified with nested PCR primers (Supplementary Methods) showing additional band in $SF3B1^{K700E}$ and absent in $SF3B1^{WT}$. Bands in the gel pictures are marked with grey and black arrows, which correspond to isoforms associated with the cryptic and canonical 3'SS respectively.

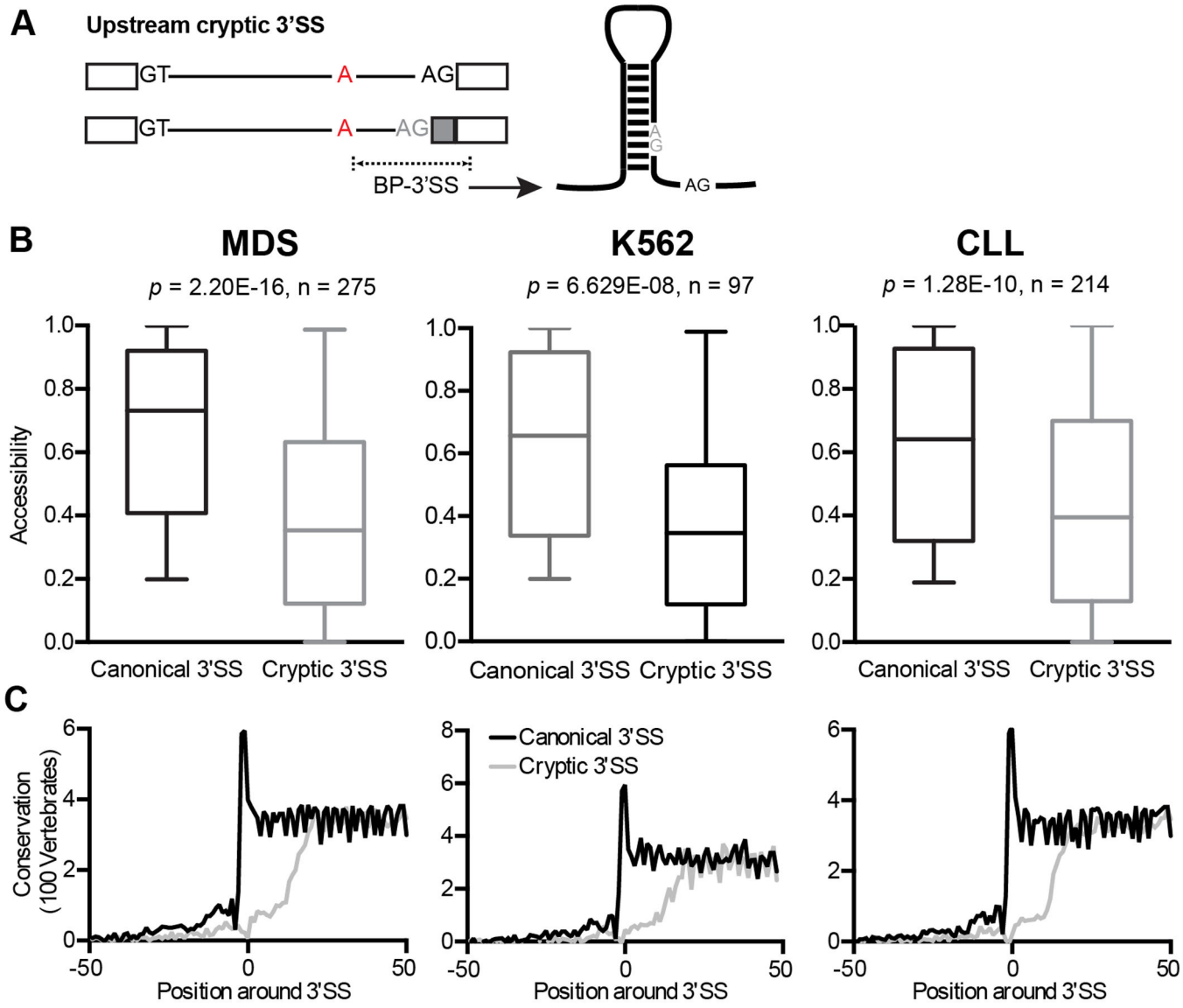

Figure 2. Cryptic 3'SS are relatively less accessible and less conserved compared to canonical 3'SS

(**A**) Schematic representation of cryptic 3'SS ('AG' in grey) located upstream of canonical 3'SS ('AG' in black) where both are associated with the same branch point ('A' in red). Exons and introns are denoted by boxes and lines, where grey box denotes extended 3' exon generated due to cryptic 3'SS usage. The sequence region denoted by dotted line folds into a secondary structure and base pairing of cryptic 3'SS (grey 'AG') renders its inaccessibility compared. (**B**) Box plots showing the accessibility values for canonical (black box) and cryptic 3'SS (grey box) for the three datasets, MDS, K562 and CLL. Each box denotes, median (middle line), upper (25%) and lower (75%) quartile with whiskers indicating maximum and minimum values. Wilcox signed-rank *p*-value denoted as '*p*' defines the statistical significance of the difference in accessibilities between 'n' number of canonical and cryptic 3'SS. Accessibility of canonical and cryptic 3'SS was determined in the sequence region between branch point (after excluding 8 nucleotides downstream of branch

point ‘A’) and 3’SS plus 15 nucleotides downstream (defined as BS-3’SS). The accessibilities were calculated in 16 windows (i.e. BS-3’SS + 0, 1, 2,….15) using RNAfold and averaged. (**C**) Base wise conservation around canonical and cryptic 3’SS analyzed in (B). Shown are the average conservation scores (−log p-values) from 100 vertebrates (phyloP100way) in the three datasets MDS, K562 and CLL.

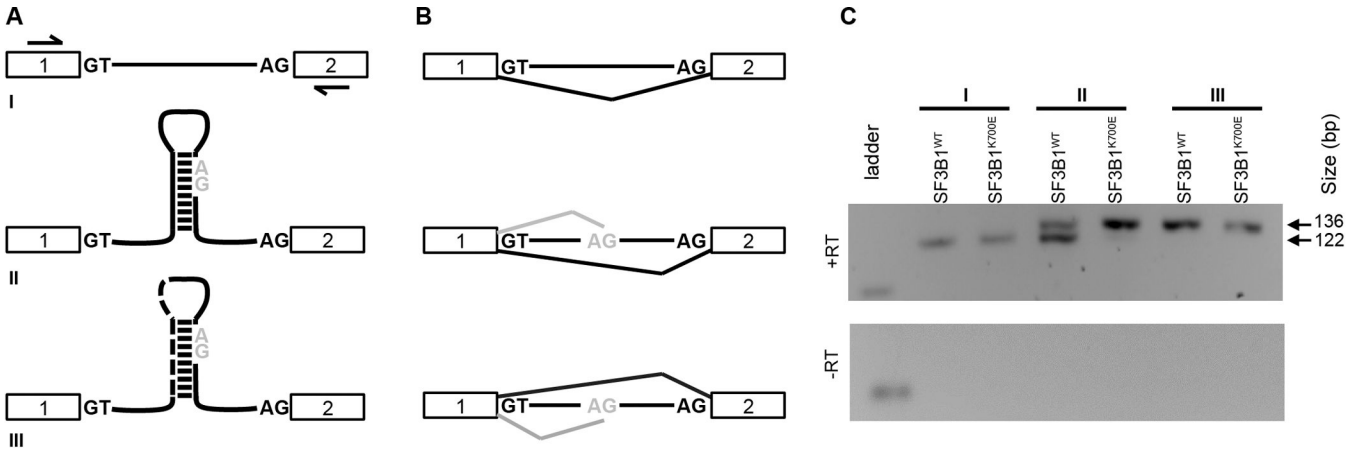

Figure 3. Mini-gene splicing assay to determine use of cryptic 3'SS within hairpin structure by SF3B1^{K700E}

(A) Scheme of mini-gene splicing constructs cloned into RG6 vector. Mini-gene I contains a control intron (91 bp long and modified from murine IgH intron) with a 5'SS, 3'SS, PPT and BP, cloned between exonic elements of chicken Troponin T. Mini-gene II incorporates a 12 base-pair hairpin upstream of the canonical 3'SS ('AG' in black) and contains a cryptic 3'SS ('AG' in grey) within the hairpin (14 bp upstream of the canonical 3'SS. Mini-gene III differs from II in being devoid of the 5' arm of the hairpin, thus no hairpin loop is predicted to be formed. **(B)** Predicted 3'SS use for each of the three constructs. Canonical AG and corresponding splicing event are in black, while the cryptic AG and splice event are shown in grey. In mini-gene I, canonical 5'SS and 3'SS use will result in a 122 bp amplicons. In mini-genes II and III, use of the downstream AG will also result in a 122 bp product, but if the upstream AG is used, the product is 136 bp (additional 14 bp long). Based on our hypothesis, upstream AG of mini-gene II is not favored in normal conditions as it is sequestered within the hairpin, hence downstream AG will be used preferentially. In mutant SF3B1-containing spliceosomes, we hypothesize that this sequestration can be overcome and hence the upstream AG will be favored (resulting in the 136 bp product). In mini-gene III, the upstream AG is favored (in both SF3B1^{WT} and SF3B1^{K700E} spliceosomes) given there is no hairpin structure. **(C)** Agarose gel electrophoresis of amplicons from cDNA amplified with nested primers. Each construct was co-transfected with either the SF3B1^{WT} or SF3B1^{K700E} construct. A 122 bp product is formed with the use of canonical 3'SS; the product is 136 bp when the upstream cryptic 3'SS is used. Mini-gene II co-transfected with SF3B1^{WT} predominantly uses the canonical AG with some use of the upstream cryptic AG. When SF3B1^{K700E} is transfected, the upstream cryptic 3' AG is almost exclusively used. Both SF3B1^{WT} and SF3B1^{K700E} preferentially use the upstream AG in mini-gene III.

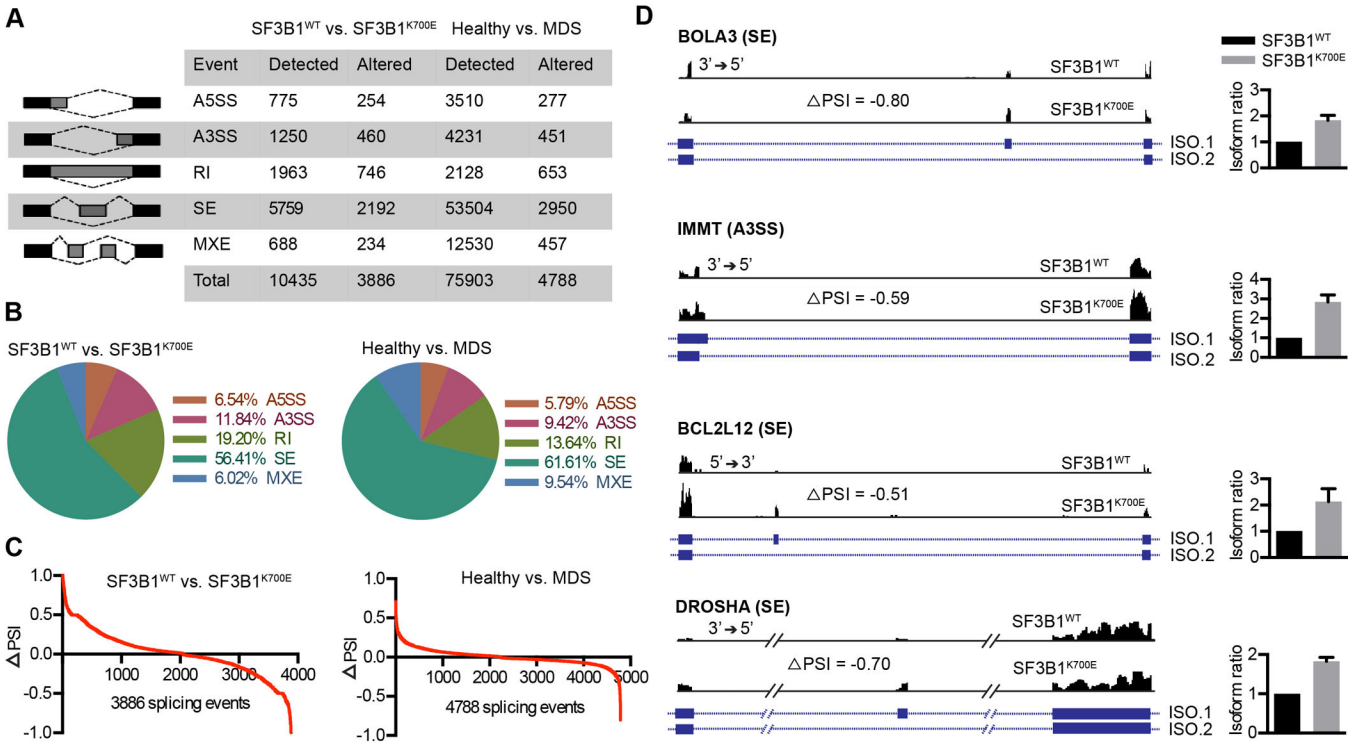

Figure 4. Altered splicing events in SF3B1^{K700E} expressing cells as analyzed by rMATS

(**A**) Number of splicing events detected and altered across different types for comparisons (between SF3B1^{WT} and SF3B1^{K700E} or CD34^{+} cells from healthy donors and SF3B1-mutant MDS patients). A5SS; alternate 5' splice site, A3SS; alternate 3' splice site, RI; retained intron, SE; skipped exon, MXE; mutually exclusive exon. (**B**) Distribution of the differentially spliced events (altered) as a percentage, showing SE (skipped exon) to be the predominant type in both datasets. (**C**) Distribution of ΔPSI of 3886 splicing events (SF3B1^{WT} vs. SF3B1^{K700E}) and 4788 events (Healthy vs. MDS patients). (**D**) Altered splicing for four candidate genes in SF3B1^{WT} compared to SF3B1^{K700E} cells. RNA-Seq coverage plots and the corresponding splicing events, ΔPSI values for the splicing isoforms (ISO.1 and ISO.2) are shown to the left of each panel. Coverage plots were normalized using Igvtool (tdf format) PCR products corresponding to ISO.1 and ISO.2 amplified from cDNA generated from SF3B1^{WT} or SF3B1^{K700E} cells and analyzed by isoforms specific qRT-PCR are shown to the right ($p < 0.05$). Three of the shown genes (BOLA3, BCL2L12 and DROSHA) have skipped exon (SE) as their primary altered splicing event while IMMT features an alternative 3' splice site (A3SS). In the rMATS analysis, difference in percent spliced in (ΔPSI) > 0 refers to higher level of exon inclusion (or longer isoform) in control sample relative to mutant, whereas ΔPSI < 0 to the inclusion level in the opposite direction. Here, exon inclusion refers to the longer isoform of the events, SE, RI, A5SS and A3SS, whereas for MXE, inclusion of first mutually exclusive exon refers to the longer isoform.

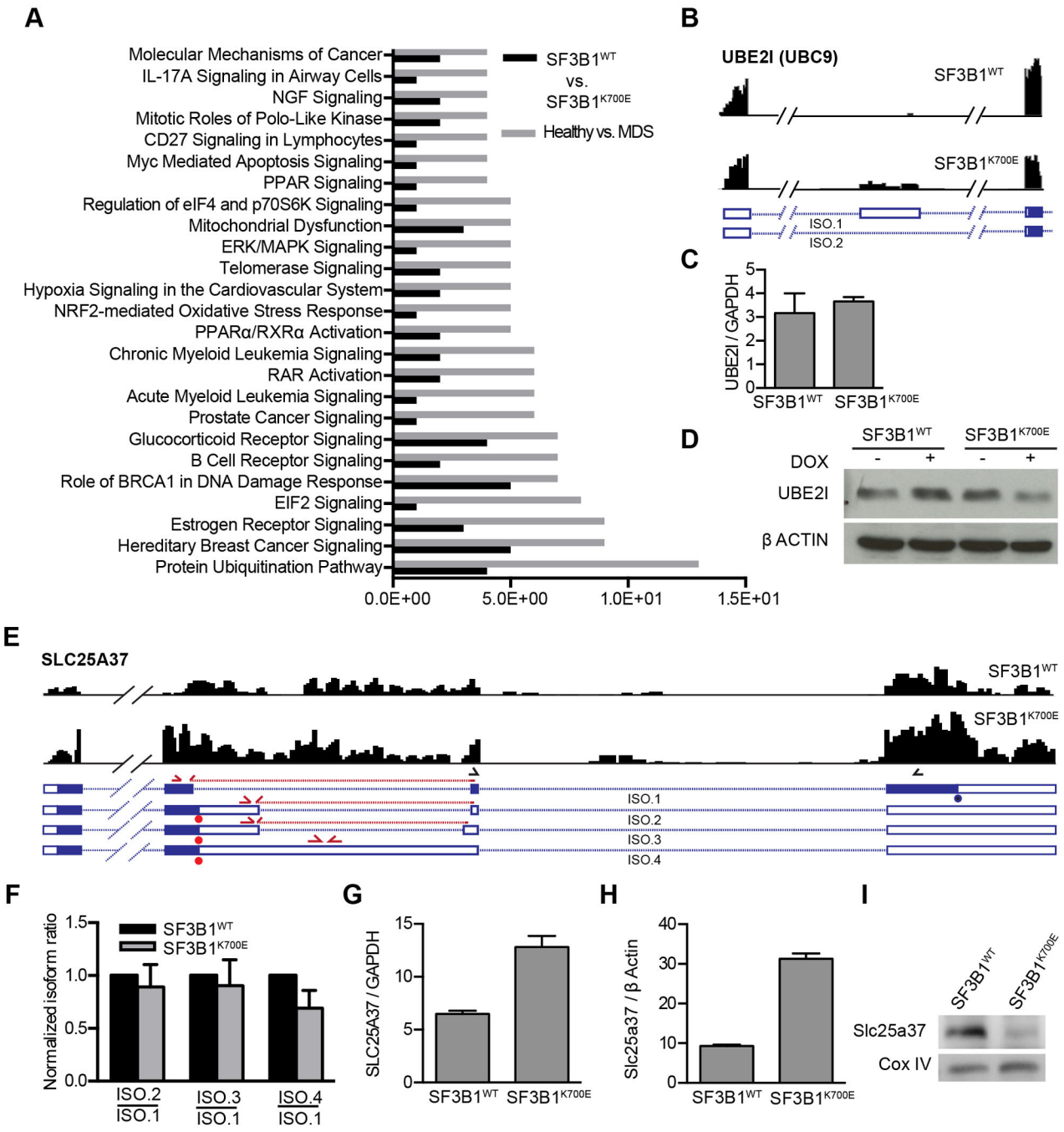

Figure 5. Differential expression of alternatively spliced genes

(**A**) Ingenuity Pathway Analysis (IPA) of genes that display significant alterations in splicing for SF3B1^{WT} vs. ^{K700E} and Healthy vs. MDS comparisons. P-values of top 25 canonical biological pathways for both datasets are shown above. Full results of IPA analysis are shown in Supplementary File 5. (**B**) Coverage plot of skipped exon (SE) in 5' UTR of UBE2I. The scale of the coverage plot is normalized using igvtools (tdf format). Exons and introns are shown as boxes and dotted lines, respectively, in blue colors. Empty box denotes 5' or 3' UTR whereas filled box denotes CDS. In the gene model, for better visualization,

long intronic region was cropped, marked as '//'. (**C**) Total transcript levels of UBE2I determined by qRT-PCR in $SF3B1^{WT}$ and $SF3B1^{K700E}$ cells. (**D**) UBE2I protein expression in $SF3B1^{WT}$ and $SF3B1^{K700E}$ cells (induced or not induced with doxycycline). $SF3B1^{K700E}$ cells show reduced expression of UBE2I. (**E**) Coverage plot (top) and the gene model (bottom) of SLC25A37 (Mitoferrin 1). Solid blue and red circles refer to authentic stop codon in isoform 1 (ISO.1) and premature termination codon (PTC) in ISO.2, ISO.3, and ISO.4 respectively. Primer pairs spanning exon-exon border used for isoform-specific quantification using qRT-PCR (shown in panel F) are indicated as red arrows. The solid black arrows denote qRT-PCR primers used to quantify total transcript level (panel G). (**F**) Quantification of SLC25A37 isoforms using qRT-PCR. The individual isoform levels were normalized to GAPDH, and then isoform ratios were calculated as indicated. Isoform ratios in $SF3B1^{K700E}$ were normalized to $SF3B1^{WT}$. (**G**) qRT-PCR for SLC25A37 confirming higher levels of SLC25A37 in $SF3B1^{K700E}$ when compared to $SF3B1^{WT}$. P-value <0.05. (**H**) qRT-PCR results performed similarly in MEL cells ($SF3B1^{WT}$ and $SF3B1^{K700E}$). (**I**) Western blot for SLC25A37 in MEL cells with $SF3B1^{WT}$ or $SF3B1^{K700E}$ along with mitochondrial loading control (CoxIV). SLC25A37 levels are reduced in $SF3B1^{K700E}$ compared to $SF3B1^{WT}$.

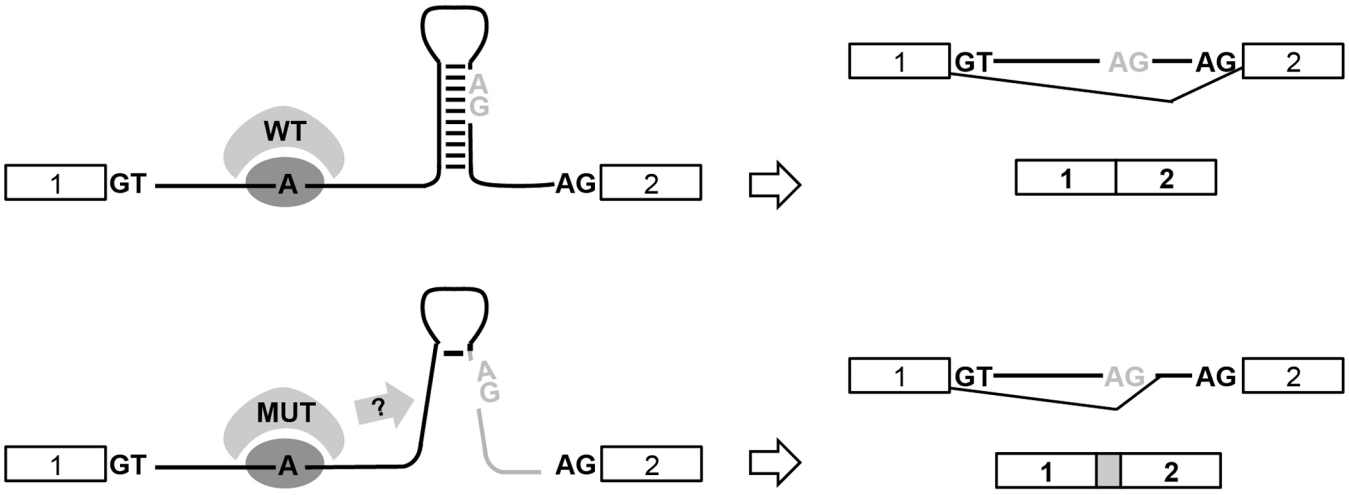

Figure 6. Proposed model for cryptic 3'SS selection in mutant SF3B1 containing spliceosomes

Cryptic 3'SS within secondary structures are inaccessible and hence not used by spliceosomes with $SF3B1^{WT}$. Canonical 3'SS usage is thus favored. With $SF3B1^{K700E}$, these cryptic 3'SS are rendered accessible by an unknown mechanism and is favored during the splicing catalysis leading to aberrantly spliced products.

Table 1

Top 50 genes with high splicing differential (ΔPSI) calculated from rMATS analysis

Shown are p-value, FDR (false discovery rate), ΔPSI and splicing event ID (includes code for splice event). A3SS; alternate 3' splice site, A5SS; alternate 5' splice site, SE; skipped exon, MXE; mutual skipped exon and RI; retained intron. Full list of the splicing events with event ID and genomic coordinates is provided in Supplementary File 2.

Gene	p-value	FDR	ΔPSI	Event ID
SREK1	2.55E-14	8.97E-13	−0.819	SE_1097
KANSL1	6.66E-16	2.45E-14	−0.778	A3SS_7248
AMBRA1	3.55E-13	1.03E-11	−0.756	SE_5506
HJURP	1.26E-10	2.02E-09	−0.743	MXE_1835
HCFC1R1	2.22E-16	9.91E-15	−0.741	SE_7490
MAFG	3.27E-05	0.000194692	−0.741	SE_33670
FANCA	2.26E-07	2.19E-06	−0.732	A5SS_3135
ZNF678	1.69E-07	1.70E-06	−0.732	SE_30560
USP8	1.05E-12	2.91E-11	−0.731	SE_38514
MTG1	6.74E-12	7.39E-11	−0.716	RI_2562
CHKA	5.05E-11	1.07E-09	−0.714	SE_33593
ACIN1	3.91E-12	9.92E-11	−0.704	SE_11916
DROSHA	6.96E-14	2.30E-12	−0.701	SE_8082
A1BG-AS1	1.42E-07	9.42E-07	−0.694	RI_5560
HTRA2	4.66E-05	0.000217474	−0.69	RI_2413
BRD9	9.26E-12	2.17E-10	−0.688	SE_19355
IP6K2	2.06E-06	1.61E-05	−0.676	SE_29795
MAP3K4	2.90E-07	2.76E-06	−0.666	SE_21142
MAP4K4	2.32E-12	5.68E-11	−0.663	A3SS_677
CHURC1-FNTB	1.30E-14	4.76E-13	−0.661	SE_2413
DGUOK	3.29E-10	5.98E-09	−0.659	SE_5760
TM7SF2	4.69E-09	3.67E-08	−0.647	RI_2828
CHURC1	4.02E-14	1.39E-12	−0.646	SE_7450
IKBKB	3.99E-05	0.000224775	−0.644	MXE_2099
TCEB1	9.61E-07	8.18E-06	−0.642	SE_36324
EYA3	7.10E-07	6.19E-06	0.851	SE_31530
NUBP2	9.19E-13	1.13E-11	0.819	RI_302
NXF1	8.86E-11	1.79E-09	0.772	SE_8919
TSPAN32	4.21E-13	5.55E-12	0.771	RI_2273
CENPA	1.35E-12	3.69E-11	0.752	SE_34124
SDCBP	8.71E-13	2.42E-11	0.734	A3SS_593
GAB1	9.51E-08	1.01E-06	0.728	SE_7059
ST6GALNAC6	1.24E-07	1.28E-06	0.726	A5SS_3525

Gene	p-value	FDR	ΔPSI	Event ID
RALGDS	4.10E-06	3.11E-05	0.72	A3SS_6485
PAN3	2.00E-05	0.000124966	0.699	SE_16872
CD97	6.57E-12	1.58E-10	0.682	SE_3142
ANKMY1	4.77E-07	4.30E-06	0.674	SE_8636
CARS	9.86E-10	1.39E-08	0.672	A5SS_3141
HMBS	9.33E-07	5.49E-06	0.665	RI_436
CDC42BPB	0.001238609	0.004610955	0.651	SE_33495
TAF1	4.31E-07	3.92E-06	0.633	SE_13153
TTC14	1.26E-13	1.80E-12	0.631	RI_2545
KDM6A	5.55E-16	2.33E-14	0.631	SE_2114
ENOPH1	2.00E-15	7.99E-14	0.63	SE_19824
NUBP2	3.58E-08	4.08E-07	0.629	A5SS_274
ZDHHC4	7.17E-06	5.01E-05	0.628	A5SS_4184
RHOT2	2.72E-07	1.74E-06	0.626	RI_471
AC004381.6	7.83E-13	2.19E-11	0.626	SE_4932
NLRP2	2.97E-05	0.000182908	0.625	A5SS_1814
VEZT	0.001241686	0.005177456	0.621	MXE_1094

Oncogene. Author manuscript; available in PMC 2017 February 24.

Table 2

High confidence genes with differential expression between K562 cells expressing SF3B1^{WT} and SF3B1^{K700E}

The significant differentially expressing genes were filtered based on at least 10 FPKM value for each of two samples. The unfiltered list of differentially expressing genes is provided in Supplementary File 5. The statistical parameters shown for the differentially expressed gene are the uncorrected p-value and the q-value defining FDR (<5%)-adjusted p-value.

Gene name	FPKM		Log_2 (fold change)	p-value	q-value
	SF3B1^{WT}	SF3B1^{K700E}			
DLK1	865.71	175.09	−2.30579	5.00E-05	0.00207154
EGR1	49.5655	220.208	2.15145	5.00E-05	0.00207154
IFITM1	1044.53	2370.26	1.18219	5.00E-05	0.00207154
LOC646719	47.9004	13.9601	−1.77872	5.00E-05	0.00207154
SLC6A6	53.1827	169.983	1.67637	5.00E-05	0.00207154
ACSM3[1]	273.759	97.5067	−1.48933	0.00015	0.0061008
INSIG1	100.507	199.311	0.987725	0.0003	0.0119561
USP5	18.8043	56.8267	1.59551	0.00035	0.0138359
SH3BGRL3	361.8	172.487	−1.06871	0.0005	0.0193789
SLC25A37[1,2]	66.2607	158.409	1.25743	0.0005	0.0193789
LONP1	159.405	80.0349	−0.993997	0.00055	0.021149
TMEM123	23.7741	60.8933	1.35689	0.0007	0.0263772
HDAC7[1,2]	75.293	31.1647	−1.2726	0.00115	0.0411948
RIF1	36.3848	18.0573	−1.01075	0.00115	0.0411948
GYPE	40.8368	133.274	1.70645	0.00125	0.0441938

These genes also show [1]differential splicing (FDR < 5%, ΔPSI ≥ 10%) between SF3B1^{WT} and SF3B1^{K700E}, and [2]NMD-eliciting feature(s) as detailed in the results section.