Splicing motifs are organized within global structural scaffold of a pre-mRNA

Kaushik Saha[1,*], Mike Minh Fernandez[1,2], Tapan Biswas[1], Simpson Joseph[1], Gourisankar Ghosh[1,*]

Department of Chemistry and Biochemistry, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0375

[2]Current address: Irell and Manella Graduate School of Biological Sciences, Beckman Research Institute of City of Hope, 1500 E Duarte Rd, Duarte, CA 91010-3000

*Correspondence: gghosh@ucsd.edu, kasaha@ucsd.edu

ABSTRACT

The specific recognition of splice signals at or near the exon-intron junctions is not explained by their weak conservation across the mammalian transcriptome. We explored the possibility of the global structural scaffold of a pre-mRNA guiding early spliceosomal components to the splice signal sequences. We find that mutation in any of the splice signal sequences could disrupt the global structure of a pre-mRNA, and consequently impede recruitment of early spliceosomal components to the pre-mRNA even when their cognate splice signals are intact. We also find distribution of pre-mRNA segments potentially interacting with early spliceosomal component U1 snRNP across the intron, spatial proximity of 5′ and 3′ splice sites within pre-mRNA scaffold, and an interplay between the global structural scaffold and splicing regulatory elements in recruiting early spliceosomal components. These results provide a conceptual base to understand the role of pre-mRNA global structural scaffold in regulating the functionalities of splicing motifs.

INTRODUCTION

The early spliceosome defines the exon-intron boundaries within pre-mRNAs for eventual removal of intron and ligation of exons by the mature spliceosome. The former assembles at the exon-intron junction by stepwise recognition of multiple splice signals on pre-mRNAs by a set of highly conserved core spliceosomal components (1). The four major splice signals, namely the 5′ and 3′ splice sites (SS) located at the 5′- and 3′-ends of the intron, the branch-point site (BS) located approximately 30 to 50 nucleotides (nt) upstream of the 3′SS, and the polypyrimidine tract (PPT) of varying lengths present between the BS and the 3′SS are bound by the U1 small nuclear ribonucleoprotein (U1 snRNP), U2 auxiliary factor 35 (U2AF35), splicing factor 1 (SF1), and U2 auxiliary factor 65 (U2AF65), respectively (1). In the early spliceosome, the 5′SS base-pairs with a nine nucleotide stretch (5′AUACΨΨACC…3′) at the 5′ end of U1 snRNA, the RNA component of U1 snRNP. Splicing could be constitutive, where the same set of splice sites are selected in all molecules of the synthesized transcript, or alternative, where more than one set of splice sites are selected in different cell or tissue types or in the same cell type under different conditions.

Splice signal recognition is a highly efficient process and errors in splice signal recognition including splicing from unannotated splice sites (cryptic splicing) often leads to diseases (2,3). How the yeast mature spliceosome prevents incorporation of a sub-optimal/cryptic splice site is currently being investigated (4,5). However, the mechanism of efficient and faithful recognition of the enormous combinatorial library of splice signals in the mammalian transcriptome by the early spliceosomal components is still not fully understood. This is primarily because the conserved regions of the mammalian splice signals are short, often about two nucleotides; this level of degeneracy within the splice signal sequences often blurs the difference between authentic splice signal sequences and random sequences to the common eye. That is why attempts at generation of an effective ‘splicing code’ that could explain the outcome of a splicing event, particularly where multiple pairs of alternatively selected splice sites are present, has remained the focus of splicing research (6,7). So far, three features of the pre-mRNA are considered to contribute to the ‘splicing

code': the sequence of the splice signals, local and long-distance pre-mRNA secondary structure, and the distribution of splicing regulatory elements (SREs) that could be exonic/intronic enhancers/silencers. SREs recruit from a large repertoire of RNA binding proteins (RBP) for managing the recognition-potential of a splice signal (8). Recently, use of deep learning technology and mathematical modeling have greatly enhanced our ability to predict splicing outcome using the genomic sequence (3,9). However, understanding of the mechanistic details are still obscure.

Given that enhancers augment the pre-mRNA features that are recognized by the early spliceosomal components, understanding the mechanism of enhancers is important for understanding splice signal recognition mechanism. The well-studied exonic splicing enhancers (ESE) bind one of numerous RBPs to promote incorporation of a specific splice signal into the early spliceosome and its functionality is negatively impacted by the level of base-pairing within the site (10). In order to identify and characterize the potential ESE-regulated splicing events (both constitutive and alternative) in the mammalian transcriptome, distribution (11-20) and local secondary structural environment (11,17) of the ESEs are intensely being investigated. With serine-arginine-rich (SR) proteins (21), one of most well-characterized ESE-recruited RBPs, it has been demonstrated that ESE-bound RBPs function to promote splicing by directly interacting with the early spliceosomal component and stabilizing the latter at the splice signal (22). Recently, it has been reported that assembly of the early spliceosome could also be promoted by structural remodeling of the pre-mRNA mediated by ESE-recruited RBPs (17). Despite significant advancement in understanding of the 'splicing code', the context-dependence of splice signal usage still cannot be mechanistically explained (23). This suggests that the magnitude and diversity of the features of the splicing code that regulate the strength of a splice signal are yet to be fully comprehended.

Biochemical and structural studies of yeast early spliceosomal complex (E-complex) suggest the formation of an early cross-intron bridge between the splice sites (24,25). In the current study, we explored if specific secondary and tertiary structural features within the global

three-dimensional structural scaffold of the protein-free pre-mRNA mediate cross-intron communication and guide the early spliceosomal components to appropriate sites. We tested this possibility by examining recruitment of early spliceosomal components to a protein-free splicing substrate *in vitro*. Recruitment of early spliceosomal components strictly dependent on ESEs might complicate the experimental analysis of the effect of the pre-mRNA structure on defining the splice signal. Therefore, we searched for a splicing substrate, where U1 snRNP recruitment is sufficiently strong and specific even in the absence of ESE-dependent functionalities, and found *AdML* (26). Mutation in any of the splice signal sequences in *AdML* disrupted its global structure and recruitment of early spliceosomal components U1 snRNP, U2AF65, and U2AF35. We also correlated the strandedness and spatial distribution of nucleotides across the protein-free pre-mRNA to their potential for interaction with U1 snRNP anchored at the 5′SS. Additionally, we find that the global pre-mRNA structure that 'integrates' the splice signals also regulate the stabilizing effect of ESE-associated RBPs on early spliceosomal components for binding to the splicing substrate.

Overall, these results suggest that the splice signals could be communicating among themselves and with ESEs within an organized pre-mRNA structural scaffold prior to assembly of the early spliceosome to distinguish an authentic splicing substrate from a random sequence.

MATERIALS AND METHODS

Cloning and protein expression

The cDNAs of all proteins except for that of full-length U2AF35 were cloned into T7 promoter-based *E. coli* expression vectors and were expressed as non-fusion or hexa-histidine fusion proteins. Proteins were expressed in *E. coli* BL21 (DE3), BL21 (DE3) pLysS, or Rosetta (DE3) cells overnight without (leaky expression) or with isopropyl β-D-1-thiogalactopyranoside induction. His_6-tagged proteins were purified by Ni^{2+}-nitrilotriacetate (Ni-NTA) chromatography while non-fusion proteins were purified by a combination of SP Sepharose (Cytiva), hydroxyapatite (Bio-Rad), and butyl Sepharose (Cytiva). His_6-tagged U2AF35 was expressed and purified from baculovirus-infected Sf9 cells as described before (27) and purified by Ni-NTA chromatography under denaturing conditions (8M urea). Then the protein was refolded by removing the urea stepwise from the buffer through multiple rounds of dialysis. All these proteins were further purified by size-exclusion (Superdex 75; Cytiva) chromatography. His_6-tagged Sm core proteins were co-expressed in combinations (D3-B, D1-D2, and E-F-G) and purified with similar procedures as described before (28). His_6-U2AF65 was either expressed on its own or co-expressed with His_6-tagged truncated SF1 (1-320 amino acids) in *E. coli*. After purification, $SF1_{1-320}$ and U2AF65 could be separated from each other by gel filtration. Full-length SRSF1 (hyperphosphorylated mimetic with all serine residues of the RS domain [197-246 a.a.] replaced with glutamate (29)) and the RNA binding domain (RBD) of SRSF5 (1-184 a.a.) are used in chromatographic and amylose pull-down assays. The RBD of SRSF1 (1-203 a.a.) was used for EMSA since the full-length SRSF1 precipitates upon contact with the native gel running buffer in the well. All SR proteins were freed from most of the *E. coli* RNA bound to it by mixing the proteins with Q Sepharose. A chimera of maltose binding protein (MBP) and bacteriophage MS2-coat protein (MS2) were expressed in *E. coli* as His_6-tagged protein and purified by Ni-NTA chromatography. The chimeric protein migrated as a doublet on SDS gel and protein(s) represented in both bands comprised of fully functional

MBP-MS2. GST-tagged proteins used for pull-down assays were purified by Glutathione Sepharose (Cytiva) and then were cleaned up further as indicated above. All purified proteins were confirmed to be RNase-free by incubating a small aliquot of the purified protein with purified U1 snRNA overnight at room temperature and analyzing the RNA quality by urea-PAGE following phenol extraction.

RNA constructs

AdML sequence was as described before (17). All pre-mRNAs except for *AdML* EH mutant contained three MS2-coat protein binding loops at their 3′ end as reported before (30). Changes in nucleotide sequences for mutagenesis of splice signals of *AdML*: Δ5′SS (31) (GTTGGGgtgag > ATTGGAaccac), ΔBS (tgctgac > tgccgtt), ΔPPT (cctgtccctttttttcc > ggagagggaaaaaaaagg), Δ3′SS (cagCT > accTC), and ΔESE (TCA > GAT). Fourteen nucleotide long *Ron* ESE sequence is as reported before (32) (UGGCGGAGGAAGCA), *AdML* 5′SS RNA (GUUGGGGUGAGUAC), and *β-globin* 5′SS (GGGCAGGUUGGUAU).

***In vitro* reconstitution and purification of U1 snRNP**

For reconstitution and purification of U1 snRNP, full-length U1 snRNA was transcribed in large scale *in vitro* using run off transcription from T7 promoter and treated with DNase I. U1 snRNP was assembled as described before (28) and purified by anion exchange chromatography (Mono Q; Cytiva) using a KCl gradient (from 250 mM KCl through 1M KCl). Particles were flash-frozen in liquid N_2 and stored at -80 ^{o}C in single-use aliquots.

***In vitro* assembly and purification of pre-mRNA complexes**

AdML IVS1 was synthesized by run-off transcription and treated with DNase I (New England Biolabs). For chromatographic purification, reconstitution of pre-mRNA complexes was carried out in 500 µl volume except for where U2AF65 and U2AF35 were added; in the latter cases, reconstitution was carried out in 1 ml volume. 100 nM pre-mRNA in complex with 3X MBP-MS2 was incubated with 300 nM U1 snRNP for 5 min at 30 °C in 20 mM HEPES-NaOH pH 7.5, 250 mM

NaCl, 2 mM $MgCl_2$, 1 mM DTT, 0.5 M urea, 0.3% poly(vinyl alcohol) (PVA, Sigma P-8136), and 5% glycerol. Then SRSF1 (650 nM) was added where indicated and incubated for 5 min at 30 °C; thereafter, U2AF65 (150 nM), $SF1_{1-320}$ (150 nM), and U2AF35 (150 nM) were added where indicated and incubated further for 5 min at 30 °C. Where SRSF5 (500 nM) is added instead of SRSF1, the SR protein is added before adding U1 snRNP. Then the reaction was incubated at room temperature for 5 min, followed by high-speed centrifugation at 4 °C for 2 min. The complexes were then bound to 1 ml Mono Q column (Cytiva) in 20 mM HEPES-NaOH pH 7.5, 250 mM NaCl, 2 mM $MgCl_2$, 1 mM DTT at 4 °C. The elution buffer had identical components as the binding buffer except for having 2M NaCl instead of 250 mM. Complexes formed with *AdML* variants were eluted with a combinatorial gradient from 0 – 20% for 5 ml followed by 20 – 30% for 10 ml. Each fraction size was 300 µl. The pre-mRNA-containing fractions were concentrated by binding to amylose resin (New England Biolabs). For SDS PAGE analysis, the resin beads were boiled in Laemmli sample buffer. For RNA isolation, the resin beads were incubated with proteinase K (New England Biolabs) in proteinase K buffer followed by phenol:chloroform extraction and ethanol precipitation. The precipitated RNA was directly dissolved in 9M urea dye for analysis by urea PAGE. We routinely observed a proportion of the pre-mRNA precipitated in the presence of U1 snRNP on the column and this proportion was slightly higher for the recruitment-defective pre-mRNAs.

For amylose pull-down assay (without chromatographic separation), the reaction was rotated with amylose resin for 20 min at room temperature followed by 5 min at 4 °C. Thereafter, the supernatant was removed by gentle centrifugation, the resin was washed with chilled binding buffer (same as binding buffer in chromatography experiments), followed by boiling in Laemmli sample buffer for SDS PAGE analysis.

Negative stain electron microscopy

The EM grid was prepared by depositing particles on the Carbon-Formvar grid (01754-F F/C 400 mesh Cu from Ted Pella) activated by 30 seconds of glow discharge, briefly washing twice with deionized water, and staining with 0.5% Uranyl Acetate for 1 min. Images were collected in FEI Tecnai G2 Sphera. 2D-average analysis was carried out using program EMAN2.

In vitro selective 2′ hydroxyl acylation followed by primer extension (SHAPE) by mutational profiling (SHAPE-MaP)

To detect both stable and dynamic interactions between the pre-mRNA and the U1 snRNP, we carried out SHAPE-MaP experiment with protein-free and U1 snRNP-bound WT *AdML* using 2 mM NMIA; low NMIA concentration was beneficial for detecting weak interactions. For carrying out SHAPE with 2 mM N-Methylisatoic anhydride (NMIA, Millipore-Sigma) with or without U1 snRNP, RNA was denatured in the presence of 5 mM EDTA pH 8.0, 50 mM NaCl at 95 °C for 3 min and then renatured by immediately placing on ice for 15 min. After neutralization of EDTA with $MgCl_2$, pre-mRNAs were incubated at 30 °C under the same conditions used for assembly of the pre-mRNA complexes as indicated above in 1 ml volume. First, the MS2-tagged pre-mRNAs were bound to 3X MBP-MS2 protein, then either U1 snRNP or U1 snRNP storage buffer (20 mM HEPES pH 7.5, 400 mM KCl, 2 mM $MgCl_2$, 15 mM Arginine-HCl, 15 mM Glutamate-KOH, 5 mM DTT) was added to the reaction mixture. After incubation, the complexes were mixed with amylose resin (New England Biolabs) and was rotated at 4 °C for 30 min. Next, the resin was separated from the supernatant by gentle centrifugation and complexes bound to the resin was eluted with 400 µl of the reaction buffer with 20 mM maltose by rotating the tubes for 1 hour at 4 °C. The pre-mRNA-containing solutions were then transferred to another tube containing 6.67 µl 120 mM NMIA solution (in DMSO) or the same volume of DMSO and mixed by pipetting immediately. Then, the tubes were kept at 16 °C for 1 hour. Finally, the RNA was purified by RNeasy Mini Kit (Qiagen) including a proteinase K digestion step following manufacturer's instructions.

For modeling secondary structure of the pre-mRNA variants, we carried out SHAPE-MaP with 8 mM NMIA, which was beneficial for generating high mutation rate. This might provide less information regarding the weak/transient interactions but will differentiate non-transient/robust interactions of different strengths, thus aiding in generation of the most stable secondary structural models. For carrying out SHAPE with 8 mM NMIA, the pre-mRNAs were not bound to any protein and the reaction buffer was same as with pre-mRNA complex formation. Denatured and renatured pre-mRNAs were incubated at 30 °C in the reaction buffer for 5 min in 1 ml volume at 100 nM concentration, then the reaction was transferred to a tube containing 66.7 µl 120 mM NMIA in DMSO or DMSO alone and incubated at 30 °C for 30 min. Then RNA was precipitated from the reaction with 300 mM sodium acetate and 3X volume of ethanol and the pellet was dissolved in 50 µl water. Then the RNA was purified by Monarch RNA cleanup kit (NEB).

The SHAPE reactivity of protein-free pre-mRNAs obtained by the two methods described above are likely to differ at places primarily because of differential reactivity of individual nucleotides at different concentrations of NMIA. Additionally, in the first case, the RNA is passed through a long processing step before it is treated with 2 mM NMIA while in the second case, the RNA is treated with 8 mM NMIA immediately after refolding; thus, alteration in certain structural features of the RNA is likely to occur in the former case before it is treated with NMIA.

Denatured control for each RNA was generated as previously described before (33) where denatured RNA was treated with 2 mM or 8 mM NMIA for 5 min at 95 °C.

Reverse transcription was carried out in MaP buffer as described before (33) and sequencing library was prepared by limited cycle PCR amplification. The samples were deep-sequenced in multiplex using MiSeq platform (Illumina) using MiSeq Nano kit (2x 250 bp, paired end) at the IGM Genomics Center, University of California, San Diego, La Jolla, CA. The demultiplexed FastQ Files obtained from the facility was processed using ShapeMapper pipeline (33) following authors' instructions to obtain the SHAPE reactivity. All negative SHAPE reactivity values were considered zero.

Two-color fluorescence measurement

RNA was denatured in 50 mM NaCl and 5 mM EDTA pH 8.0 at 95 °C for 3 min and then placed on ice. Then 200 nM RNA was mixed with 200 nM of DNA probes, heated at 55 °C for 5 min, and then gradually cooled to room temperature over a period of 60 min. This allowed the DNA probes to anneal to the loop areas of the folded RNA. Three samples were prepared for each RNA: one containing Cy3-labeled 5′ probe and Cy5-labeled 3′ probe, one containing Cy3-labeled 5′ probe and unlabeled 3′ probe, and one containing unlabeled 5′ probe and Cy5-labeled 3′ probe. Next, EDTA was neutralized with $MgCl_2$. Thereafter, HEPES pH 7.5, NaCl, $MgCl_2$, and PVA were added to the final concentration of 20 mM, 250 mM, 2 mM, and 0.3%, respectively, and the reaction was incubated at 30 °C for 5 min. Fluorescence was measured in Jasco Spectrofluorometer 8500 with a bandwidth of 5 nm for excitation and emission. All measurements with intact *AdML* were replicated thrice from distinct samples. For annealing of the structure-disrupting 40-nt long DNA to *AdML* RNA, 200 nM RNA, 200 nM disruptor DNA, and 200 nM labeled DNAs were mixed in the presence of 50 mM NaCl and 5 mM EDTA pH 8.0 and heated at 95 °C for 3 min and then was gradually cooled for over 60 min to room temperature. Thereafter, the RNA is treated similarly as folded RNA.

Electrophoretic mobility shift assay (EMSA)

EMSA was carried out as described before (17).

Transfection-based splicing assay

Splicing assay was carried out as described before (17). Antisense morpholino oligonucleotide (AMO) for the 25-nt region at the 5′ end of U1 snRNA (U1 AMO) was as described before (34): 5′-GGTATCTCCCCTGCCAGGTAAGTAT-3′, scrambled AMO: CCTCTTACCTCAGTTACAATTTATA.

GST pull-down assay

GST pull-down assay was carried out as described before (29).

RESULTS:

AdML splice signal mutants with intact 5′SS are defective in U1 snRNP recruitment

For testing functional U1 snRNP binding to *AdML*, purified recombinant U1 snRNP particle (see Supplementary Figures S1A for Coomassie blue-stained SDS PAGE analysis, S1B for silver-stained SDS PAGE analysis, and S1C for negative-stained electron micrograph) was mixed with WT *AdML* (26) (in chimera with 3X MS2-protein binding loops) and MBP-MS2 fusion-protein as described before (35), and the resulting complexes were purified by anion-exchange chromatography. The chromatogram (Figure 1A – blue line) revealed one distinct and two overlapping peaks. SDS (Figure 1B) and urea (Supplementary Figure S1D) PAGE analyses of the peak fractions showing protein and RNA, respectively, indicate that the first peak (fractions 19-21) contain free U1 snRNP, the second peak (fractions 24-29) U1 snRNP:pre-mRNA complex (MBP-MS2 represents pre-mRNA but at 3 times the pre-mRNA level) and the third peak (fractions 30-32) diminishing level of U1 snRNP and high level of MBP-MS2 suggesting trailing of the U1 snRNP: pre-mRNA complex overlapping with excess free pre-mRNA. The fractions indicated as “pre-mRNA complexes” (fractions 24-32 in Figure 1B) were concentrated by MBP-MS2-pull-down using amylose resin before PAGE analysis.

Next, we examined U1 snRNP binding to the splice signal mutants of *AdML*. Similar experiments with 3X MS2-tagged *AdML* mutants (Δ5′SS, ΔBS, Δ3′SS, ΔPPT) (Figure 1A) indicate that the level of U1 snRNP co-eluted with *AdML* mutants is distinctively less than that for the WT substrate as reflected in the SDS-PAGE analyses of the peak fractions (Figures 1C, 1D, 1E, 1F). A representative urea gel demonstrating the total RNA content in the fractions (pre-mRNA and U1 snRNA) is also shown for *AdML* Δ5′SS (Supplementary Figure S1E) suggesting significantly weaker co-elution of U1 snRNA with *AdML* Δ5′SS compared to *AdML* WT. The chromatograms for *AdML* Δ3′SS and ΔBS exhibited abolition of the second peak (Figure 1A) while those for *AdML* Δ5′SS and ΔPPT exhibited a shoulder in place of the second peak. The shoulder likely represents certain conformational state(s) of the mutant pre-mRNAs.

The importance of 5′ and 3′ exonic segments in U1 snRNP recruitment was also examined by the chromatographic U1 snRNP binding assay. Earlier literature indicates that loops immediately upstream of the 5′SS could be important for early spliceosome assembly (17). We assessed involvement of these loops in recruiting U1 snRNP by examining binding of U1 snRNP to the splicing-defective exonic hybridization (EH) mutant (17). This mutant contains several altered nucleotides (spanning positions 29 to 55, as numbered in Figure 2A) upstream of the 5′SS (17) to eliminate loops in this region. Comparative RNA profile analyses in chromatographic fractions (Supplementary Figure S1F) of WT vs *AdML* mutant complexes shows a markedly reduced co-elution of U1 snRNA with *AdML* EH mutant (Supplementary Figure S1G) suggesting their weak interaction. We also performed similar analysis with an *AdML* construct lacking the 3′ exon (*AdML* ΔEx2) to examine if the 3′ exon participates in stabilizing U1 snRNP (Supplementary Figure S1F). The level of U1 snRNA co-eluted with *AdML* ΔEx2 (Supplementary Figure S1H) was distinctively less when compared to that of the WT *AdML* (Supplementary Figure S1D).

These data indicating impairment of U1 snRNP recruitment upon mutation or truncation far away from the 5′SS strongly suggest an important role of the global structural landscape of *AdML* pre-mRNA in U1 snRNP recruitment.

Mutation in major splice signals disrupts global three-dimensional structure of *AdML*

To understand why the splice signal mutants exhibited a weaker U1 snRNP recruitment efficiency compared to the WT substrate even when 5′SS was intact, we hypothesized that mutation in the splice signal sequences disrupt the three-dimensional structural scaffold of *AdML*. Since difference in SHAPE reactivities of the WT and the mutants would be reflective of their different/altered three-dimensional structures (secondary and tertiary), we measured SHAPE reactivity of *AdML* WT, Δ5′SS, Δ3′SS, ΔBS with SHAPE-MaP technique (33) with 8 mM NMIA and derived their secondary structure models with RNAstructure (36). Each of the *AdML* splice signal mutants exhibits a different SHAPE reactivity and different secondary structure model compared to those of the WT

substrate (Figure 2A) – *AdML* Δ5′SS (Figure 2B), *AdML* Δ3′SS (Figure 2C), and *AdML* ΔBS (Supplementary Figure S2). Difference in secondary structure models also conforms to difference in their tertiary structure. The raw SHAPE reactivity values are provided in Supplementary File 1.

Overall, this result suggests that the major splice signals are functionally and structurally integrated into the global structural scaffold of *AdML* pre-mRNA.

Strandedness of nucleotides across the pre-mRNA correlates to U1 snRNP recruitment

We hypothesize that in order for the global three-dimensional structural scaffold of *AdML* to regulate U1 snRNP recruitment, U1 snRNP could potentially contact various nucleotides across the pre-mRNA, which are spatially proximal to the 5′SS region. Therefore, we examined potential points of contact between U1 snRNP and *AdML* by SHAPE experiments with 2 mM NMIA (see Methods). Then SHAPE reactivity differential of U1 snRNP-bound and protein-free (mock-treated) *AdML* ([*AdML*+U1] – *AdML*) (Figure 3A – blue line) was calculated (raw SHAPE values provided in Supplementary File 1 and plotted in Supplementary Figure S3A). Lower SHAPE reactivity of a nucleotide in the U1 snRNP-bound *AdML* compared to protein-free WT *AdML* indicates its protection, caused by direct contact with U1 snRNP components or base-pairing introduced by pre-mRNA structural remodeling. On the other hand, a higher SHAPE reactivity in the U1 snRNP-bound *AdML* means removal of structural constraints through pre-mRNA structural remodeling. We observed changes in SHAPE reactivity across the entire pre-mRNA upon U1 snRNP binding to *AdML*: distinct enhancement of flexibility was observed upstream of the 5′SS, in the middle of the intron at a terminal loop area (110-127 nt), around the BS region, immediately upstream of the 3′SS, and at two places in the 3′ exon (Figure 3A); enhanced protection was observed at various places in the 5′ exon, at and immediately downstream of the 5′SS, slightly upstream of the BS (around nt 140), and immediately downstream of the 3′SS (Figure 3A).

Since generally splicing factors are reported to interact with single-stranded RNA segments (10,17), we examined if the *AdML* mutants that fail to bind U1 snRNP efficiently have some of the

potential U1 snRNP contact points occluded by base-pairing. We estimated the SHAPE reactivity of protein-free *AdML* Δ5′SS and *AdML* Δ3′SS treated with 2 mM NMIA (raw SHAPE data shown in Supplementary File 1 and plotted in Supplementary Figures S3B and S3C). Then we calculated the following SHAPE reactivity differentials: (*AdML* Δ5′SS – *AdML* WT) (orange line of Figure 3A) and (*AdML* Δ3′SS – *AdML* WT) (grey line of Figure 3B), which are overlaid onto the plots of ([*AdML*+U1] – *AdML* WT) differential (blue line in Figures 3A and 3B). Interestingly, both (*AdML* Δ5′SS – *AdML* WT) and (*AdML* Δ3′SS – *AdML* WT) differentials exhibited low negative values at several common nucleotide positions with ([*AdML*+U1] – *AdML* WT), marked with rectangular brackets above the plots.

Overall, these results suggest that availability of certain single-stranded nucleotides across the pre-mRNA correlates to U1 snRNP recruitment efficiency to *AdML* variants.

Close spatial positioning of splice signals to 5′SS

For some of the nucleotides across the pre-mRNA exhibiting reduction in SHAPE reactivity upon U1 snRNP engagement to contact U1 snRNP components, these will have to be spatially close to each other and close to the 5′SS. Therefore, we examined if the pre-mRNA helices are close to each other in three-dimensional space in the protein-free state by Förster resonance energy transfer (FRET). We annealed a Cy3-labeled short DNA (20-nt long DNA #2) to the terminal loop area in the 5′ exon (around nt-50 of helix A) and a Cy5-labeled short DNA to either the intronic stem-loop around nt-115 of helix B (20-nt long DNA #5), the intronic stem-loop around nt-155 of helix C (19-nt long DNA #3), or the 3′ exonic stem-loop around nt-210 of helix D (16-nt long DNA #6) (Figure 4A) to *AdML*. In order to avoid major disruption of global structure of the pre-mRNA, *AdML* is first refolded and then the primers are annealed (see Methods). Since disruption of the base-pairing immediately upstream of the 5′SS is important for splicing (17), perturbation of certain base-pairs immediately upstream of the 5′SS due to annealing of the Cy3-labeled DNA #2 is not likely to affect the relevant pre-mRNA structure. For the other terminal loops, the labeled DNAs

anneal to largely single-stranded regions with minimal perturbation of local structure. Steady-state fluorescence was monitored by exciting at 550 nm and scanning for fluorescence-emission intensity at wavelengths between 560 and 720 nm (Supplementary Figures S4A, B, C). Cy3 emission (maximum at 565 nm) was reduced in the presence of Cy5 probe with all three pairs of DNA probes (2+3, 2+5, 2+6) suggesting FRET. A corresponding gain of Cy5 emission (emission maximum 665 nm) was also noted in the presence of Cy3. The level of reduction in Cy3 emission upon addition of the Cy5 probe was significantly higher with the DNA probe-pairs 2-3 and 2-6 than 2-5 (Figure 4B) suggesting that annealed DNA #2 (helix A) is further from annealed DNA #5 (helix B) than annealed DNAs #3 and #6 (helices C & D). This is reflective of early-stage proximity of 5′SS (located in helix A) with BS (in helix C) and 3′SS (in helix D) in protein-free *AdML*. We also examined effects on FRET of all three probe pairs upon disruption of *AdML* global structure with annealing of a long DNA (red line in Figure 4A). We observed no significant FRET with any of the probe pairs upon disruption (Figure 4B, Supplementary Figure S4D, S4E, S4F). The raw fluorescence values are provided in Supplementary File 2.

These results suggest that splice signals residing in different helices of *AdML* could come close to each other in the protein-free state for recruiting the early spliceosomal components.

The ESE-dependent RBP SRSF1 stabilizes U1 snRNP on WT *AdML* but not *AdML* Δ3′SS

For mammalian pre-mRNAs, functional recruitment of the early spliceosomal components at the splice signals are reported to require interactions with ESE-bound RBPs. Since we observed stable and specific U1 snRNP recruitment to protein-free *AdML*, we hypothesize that the ESE-dependent functionalities of RBPs could stabilize U1 snRNP further for preparing the complex for the next step of the assembly. Therefore, we tested the stability of U1 snRNP:*AdML* complex in the presence as well as the absence of ESE-dependent SR protein SRSF1 – known to stabilize U1 snRNP on the pre-mRNA by direct interaction (22,29,37) – by EMSA. Supplementary Figure S5A shows the SR protein constructs used in this study, all of which are reported to be functional. The truncated

SRSF1 containing only the RNA binding domain (RBD) is referred to as SRSF1-RBD. SRSF5 construct used represents its RBD. *AdML* forms complexes with U1 snRNP both in the presence as well as the absence of SRSF1-RBD (Figure 5A – lanes 3 & 6). The *AdML*:U1 snRNP complex withstood the challenge of an excess of unlabeled 14-nt long *AdML* 5′SS RNA in the presence of SRSF1 better than in the latter's absence as indicated by the release of free radiolabeled probe upon the challenge (compare lanes 4 and 7). Challenging the *AdML* complexes with 14-nt long 5′SS of *β-globin* pre-mRNA (38) released less free probe than *AdML* 5′SS RNA (compare lanes 7 & 8). Complexes formed with *AdML* Δ5′SS pre-mRNA was reduced significantly compared to that with the WT substrate in all cases (lanes 11-16) signifying a weaker interaction. Then we tested if the presence of SRSF1 could enhance U1 snRNP binding to one of the *AdML* mutants with intact 5′SS. Chromatographic purification of the mixture of *AdML* Δ3′SS, SRSF1, and U1 snRNP (Supplementary Figure 5B) and subsequent SDS PAGE analysis of the fractions containing the pre-mRNA complexes (Figure 5C) suggest that U1 snRNP binding to *AdML* Δ3′SS did not improve up to the extent of *AdML* WT (compare Figures 1B, 1E, and 5C).

Overall, these results suggest that pre-mRNA global structure dictates the functionality of the ESE-dependent RBP SRSF1 in stabilizing early spliceosomal components on the pre-mRNA.

***AdML* Δ3′SS recruits U1 snRNP in the presence of SRSF5 but not in protein-free state**

Although *AdML* Δ3′SS could not recruit U1 snRNP in the protein-free state or in the presence of SRSF1, transfection-based splicing assay indicates that *AdML* Δ3′SS splices as efficiently as *AdML* WT (Figure 5D – lanes 1 & 2) using a cryptic 3′SS (AG) that is 6-nucleotide downstream of the authentic 3′SS (Supplementary Figure S5B). Among the other splice signal mutants, *AdML* Δ5′SS and *AdML* ΔPPT exhibit abolished splicing while *AdML* ΔBS weakly splices into WT mRNA (determined by Sanger sequencing of the mRNA), likely using a different branchsite (Supplementary Figure S5C) (39). Then we examined if U1 snRNP is required for splicing of both *AdML* WT and *AdML* Δ3′SS since U1 snRNP-independent splicing has also been reported (40).

We expressed these variants of *AdML* in *HeLa* cells co-transfected with 25-nt long antisense morpholino oligonucleotide complementary to the 25-nt of the 5′ end of U1 snRNA as described before (34). Inhibition of U1 snRNP allowed accumulation of the pre-mRNA variants (Supplementary Figure S5D – compare lanes 1, 2 with 6, 7), which otherwise are quickly spliced into mRNAs. *AdML* Δ3′SS splicing being dependent on the presence of functional U1 snRNP *in vivo* suggests that U1 snRNP binding efficiency to this variant is restored *in vivo*.

The defect in U1 snRNP recruitment to *AdML* Δ3′SS could largely be attributed to unfavorable global three-dimensional structural scaffold of *AdML* Δ3′SS and could not be corrected by SRSF1. Here we examined if another ESE-dependent RBP could correct U1 snRNP recruitment deficiency in *AdML* Δ3′SS. In order to identify the ESE, we mutated three nucleotides in a single-stranded region in the 5′ exon of *AdML* Δ3′SS (Supplementary Figure S5E, the mutant is termed ΔESE), which did not significantly alter splicing of *AdML* ΔESE with intact 3′SS but drastically reduced that of *AdML* Δ3′SS ΔESE (Figure 5D – lanes 3 & 4). Blocking of U1 snRNP with U1 antisense morpholino oligonucleotide suppresses splicing of both *AdML* ΔESE and *AdML* ΔESE Δ3′SS (Supplementary Figure S5D – compare lanes 3, 4 with 8, 9) suggesting that both these splicing events are U1 snRNP-dependent.

ESE-finder (41) indicates that the region mutated in *AdML* Δ3′SS ΔESE (UCACUCU) is an *in vitro*-selected binding site for SR protein SRSF5 (alias SRp40). Therefore, we mixed *AdML* Δ3′SS with SRSF5 and U1 snRNP, and purified the complex by ion-exchange chromatography (Figure 5E). SDS PAGE indicates that high level of U1 snRNP is co-eluted with *AdML* Δ3′SS that is also bound by SRSF5 (Figure 5F). We also mixed the same components with *AdML* Δ3′SS ΔESE and purified the complexes (Figure 5E). SDS PAGE of the fractions indicate that a lower level of U1 snRNP is co-eluted with the pre-mRNA (compare Figures 5F and 5G). Slight reduction in the level of MBP-MS2 was also observed with *AdML* Δ3′SS ΔESE compared to *AdML* Δ3′SS; this is due to precipitation of a portion of the pre-mRNA complexes on the column, which was observed in case of the recruitment-defective pre-mRNAs throughout this study. For example, the level of

eluted MBP-MS2 with WT *AdML*-MS2 in the presence of U1 snRNP (Figure 1B) is slightly greater than that with MS2-tagged *AdML* mutants (Figures 1C, D, E, F). To test if SRSF5 stabilizes U1 snRNP primarily through direct interaction, we examined interaction between GST-SRSF5 and U1 snRNP by GST pull down assay (Supplementary Figure S5F). Weak interaction between U1 snRNP and GST-SRSF1-RBD was observed but none was detected between GST-SRSF5 and U1 snRNP. We also used another variant of U1 snRNP assembled with a truncated Sm B (1-174 a.a.) in the GST pull-down assay (Figure 5H), which exhibited a stronger interaction between SRSF1-RBD and U1 snRNP but none between SRSF5-RBD and U1 snRNP.

Overall, these results suggest that SRSF5 stabilizes U1 snRNP on *AdML* Δ3′SS without a detectable interaction with U1 snRNP likely through pre-mRNA structural remodeling.

Interplay between ESE and global structure of *AdML* variants regulates recruitment of U2AF65/U2AF35

Next, we examined if binding of the 3′ end recognizing factors U2AF65 and U2AF35 is affected by the above-described splice signal mutations. We mixed U1 snRNP, SF1, U2AF65, and U2AF35 with *AdML* WT (bound to MBP-MS2) and purified by anion-exchange chromatography (red line in Supplementary Figure S6A). From SDS PAGE analysis of the fractions, we did not observe a clear binding of any splicing factor other than U1 snRNP (Supplementary Figure S6B), perhaps due to requirements of additional ESE-dependent splicing factors e.g., SR proteins. Therefore, we added SRSF1 to the reaction mixture and purified the complexes by chromatography (blue line in Supplementary Figure S6A). SDS PAGE analysis of the fractions indicate a high level of U1 snRNP co-eluting with the MBP-MS2-bound pre-mRNA (Supplementary Figure S6C; raw SDS gel images are provided in Supplementary File 3); a closer examination of the gel image indicates small (sub-stoichiometric) levels of U2AF65 and U2AF35 but not SF1. In Supplementary Figure S6C and the subsequent SRSF1-containing gels, SRSF1 band intensity appears higher than other protein bands since SRSF1 stains about five-times more intensely than U1-A as well as U1-70k (Supplementary Figure S6I).

Two evolutionarily conserved serine residues (S80/S82) are heavily phosphorylated *in vivo*, which is reported to enhance interactions of SF1 to the pre-mRNA and U2AF65 (42). Therefore, we used phosphomimetic variant of SF1 – SF1(S80E/S82E) in the chromatographic purification assay (Figure 6A – blue line) to test if this enhances binding of SF1, U2AF65, and U2AF35 and noticed no enhancement of binding of U2AF65, U2AF35, and SF1 compared to when we used unphosphomimetic SF1 (Figures 6B and Supplementary Figure S6C). Next, we examined formation of this complex with the *AdML* ΔPPT (Figure 6A – grey line) followed by SDS PAGE analysis (Figure 6C) to test if this complex is specific. We reproducibly observed absence of the U2AF65 and U2AF35 protein bands. Although the presence of SRSF1 enhanced bound U1 snRNP level to *AdML* ΔPPT compared to protein-free *AdML* ΔPPT (Figure 1F), it is not as robust as observed with *AdML* WT; a peak of U1 snRNP proteins co-eluted with *AdML* WT complexes (a gradual increase between Fr 25 – 27 and decrease between Fr 28-33 in Figure 6B) is not observed with *AdML* ΔPPT. We then tested assembly of this complex with *AdML* Δ5′SS (Figures 6A – red line), which behaved like *AdML* ΔPPT (Figure 6D). We also tested recruitment of U2AF65 and U2AF35 to *AdML* Δ3′SS in the presence of SRSF5 by chromatography (Figure 6E) followed by SDS PAGE of the fractions (Figure 6F), which indicates WT-level recruitment of U2AF65 and U2AF35. We have not examined the identity of the SRSF5-enhancer, likely near the 3′SS, which promotes recruitment of U2AF65 and U2AF35 to *AdML* Δ3′SS.

Since we did not detect SF1 in the purified *AdML* WT complexes, we asked if SF1 is required for binding of U2AF65 and U2AF35 to the WT pre-mRNA. We carried out similar chromatographic experiment with *AdML* WT with all components except for SF1 (Supplementary Figure S6D). Co-elution of U2AF65 or U2AF35 with the pre-mRNA (Supplementary Figure S6E) was not as clear as was in the presence of SF1 (Supplementary Figure S6C). Additionally, neither chromatographic (Supplementary Figures S6F, S6G) nor amylose-pull-down assay (Supplementary Figure S6H) could detect SF1 or SF1(S80E/S82E) bound to the pre-mRNA in the absence of U2AF65 and U2AF35. These results suggest that in our simplified system consisting of

purified components, SF1 is required for stabilization of U2AF65 and U2AF35 on *AdML* but is not retained in the complex suggesting it may have a chaperone-like functionality. A previous report indicates that SF1 may even be skipped for several mammalian splicing events in splicing-competent nuclear extract containing all nuclear proteins (43).

Overall, mutation in non-cognate splice signals could impede ESE-dependent recruitment of U2AF65 and U2AF35 to *AdML* variants likely through disruption of global RNA structural environment.

DISCUSSION

The efficiency of a splicing event depends not only on the nucleotide sequence of the splicing motifs (i.e. splice signals and SREs) but also their context (23). Due to this context-dependent functionality of splicing motifs, the regulatory potential of a splicing factor is not straight-forwardly correlated to the level of conservation within its cognate binding motif. In the current work, we examined the potential of the global structural scaffold of a pre-mRNA to regulate the engagement of the early spliceosomal components and the functionalities of the RBPs, which bind the splice signals and SREs, respectively. Our work suggests that global structural scaffold of the pre-mRNA provides context to the splicing motifs by integrating/embedding them into a specific structured splicing unit. This structural framework also excludes the pseudo splicing motifs disallowing their recognition.

Our observations regarding the alteration in the global structure of the pre-mRNA upon mutation of any of the splice signals and the spatial proximity of the splice signals in the WT substrate suggest the integration of the splice signals into the global pre-mRNA structure. Furthermore, the defect observed in recruitment of U1 snRNP, U2AF65, and U2AF35 upon mutation of their non-cognate splice signals is suggestive of an early cross-intron communication.

We also observed difference in mechanism of recruitment of U1 snRNP for authentic and cryptic splicing of *AdML* and its mutants, respectively. A*dML* WT engaged U1 snRNP strongly for its eventual stabilization by the ESE-dependent RBP SRSF1. In contrast, *AdML* Δ3′SS, which splices efficiently from a cryptic 3′SS 6-nt downstream of the authentic 3′SS, does not recruit U1 snRNP in protein-free state or in the presence of SRSF1. This is suggestive of a global structural scaffold of *AdML* Δ3′SS that is unfavorable to engage just U1 snRNP as well as U1 snRNP with ESE-bound SRSF1. SRSF1-mediated pre-mRNA structural remodeling was also not sufficient to mediate U1 snRNP recruitment further suggesting the importance of the global structural scaffold of the pre-mRNA. The defect in recruitment of U1 snRNP to *AdML* Δ3′SS is rescued by SRSF5 when its ESE is present in the 5′ exon. We did not observe a detectable interaction between

SRSF5 and U1 snRNP. Thus, structural remodeling of *AdML* Δ3′SS by SRSF5 but not SRSF1 could have helped stabilize U1 snRNP. This agrees with our reported observation that SR proteins could bind the pre-mRNA in multiple copies in an ESE-dependent manner and structurally remodel the pre-mRNA to promote early spliceosome assembly (17). Nonetheless, a weak pre-mRNA-dependent interaction between SRSF5 and U1 snRNP in the *AdML* Δ3′SS complex cannot be ruled out.

Based on our results, we propose the possibility of two modes of cooperation between ESE-dependent RBPs and global pre-mRNA structure for recognition of a splice signal (Figure 7). The pre-mRNA with a favorable global structure could engage U1 snRNP (or other early spliceosomal components), with further stabilization by ESE-bound RBPs for assembly of the early spliceosome. The strength of interaction between the pre-mRNA and the early spliceosomal components could be regulated by pre-mRNA sequence, structure, SREs, RBP concentrations, and other environmental and contextual factors. The pre-mRNA scaffold could spatially orientate/position early spliceosomal components and the ESEs to mediate the interaction between the RBP and the early spliceosomal component. Alternately, the pre-mRNA structure may initially be not favorable to recruit U1 snRNP or other early spliceosomal components and could do so only upon structural remodeling by RBPs in an ESE-dependent manner. Both possibilities could exist in most recognition events but to different extent. In both cases, the nucleotides recognized by U1 snRNP (or other early spliceosomal components) across the pre-mRNA likely is not well-conserved like the other splice signal sequences and defined by a set of closely related structures.

Spatial proximity of the 5′SS and the BS is also observed in protein-free pre-mRNA molecules from yeast by single-molecule FRET although U1 snRNP binding to the 5′SS competes away the stem-loop containing the BS in the early spliceosome (44). It is possible that the large size of yeast U1 snRNP (~ 800 kDa) compared to human U1 snRNP (~ 220 kDa) could have pushed away BS from the 5′SS disallowing FRET. It is also possible that highly conserved splice signal sequences of yeast pre-mRNA provide a strong and specific binding to U1 snRNP (and

other splicing factors) without the involvement of additional sequences across the intron. Further investigation is required to understand if the folding state of the yeast pre-mRNA represents an ancestral trait of the mammalian pre-mRNA and is important for recruitment of splicing factors.

In vivo the splice site recognition mechanism is complex. The order of splicing of introns of highly varied lengths may not follow the order of intron synthesis (45) involving co-transcriptional or post-transcriptional splicing (46). Our model presumes that both 5′ and 3′SS are present in the pre-mRNA during substrate definition, which could be true for post-transcriptional splicing but not co-transcriptional splicing. Due to the complexity of the spliceosomal system *in vivo*, it may be difficult to put forth a universal mechanism for recognition of splice sites across the transcriptome. However, our data and its extrapolations support a model that correlates to several previously described molecular recognition events related to co-transcriptional splicing. First, context-dependent binding of splicing factors to full-length pre-mRNAs observed *in vitro* is likely correlated to their *in vivo* binding (47). Additionally, many splicing factors, including U1 snRNP, are found to be associated with RNA polymerase II (RNAPII) (48,49). U1 snRNP particle seems to be associated with both the 5′SS and elongating RNAPII molecule at least until the following exon synthesis is completed (48). This might provide U1 snRNP the opportunity to recognize the splicing substrate involving both splice sites across an intron (or an exon). Looping out of the longer introns, which brings the splice sites closer, is also reported to enhance splicing (50) possibly by allowing close-knit association of splice signals and flanking segments within a three-dimensional structural framework of pre-mRNA for functional recruitment of the U1 snRNP. The physical separation of exons and introns between nuclear speckles and nucleoplasm and the presence of U snRNPs at the periphery of nuclear speckle (51) could support cross-intron and/or cross-exon recognition of the pre-mRNA by U1 snRNP.

There are many additional questions that need to be answered to better appreciate the impact of global structural scaffold of the pre-mRNA on splicing. Both exon and intron definition of a pre-mRNA is important to explain the fidelity of splicing (48) and the early spliceosome could

assemble across the intron as well as across the exon (24). Additionally, earlier observations suggest that the polypyrimidine-tract binding protein (PTB) prevents intron-definition but does not block U1 snRNP recruitment at the upstream 5′SS (52). Thus, it would be interesting to examine if the structural framework recognized by U1 snRNP may be formed across the internal exons and how PTB and other splicing suppressors interact with pre-mRNA global structural landscape. Furthermore, a previous study of a substrate with two alternative 5′SSs suggested that both 5′SSs are engaged by U1 snRNP in the early spliceosome (E-complex) and one is purged during ATP-dependent transition to A-complex (53). Understanding if the pre-mRNA global structure could mediate such alternative splice site selection at the early stage will require further investigation. Moreover, transcriptome-wide binding studies suggest that human U2AF65 avoid the majority of non-authentic PPT-like sequences (54). Understanding how this is achieved by a coordinated action of DEK (55), hnRNP A1 (56), the global RNA structural scaffold, and other PPT-like sequence binding proteins that might occlude the unauthentic sites will require further investigation. Additionally, recruitment of U2AF65 at alternative PPTs within the same intron (54), at the authentic PPT in coordination with other PPT-binding protein factors (57), or at the authentic PPT with the help of multiple exonic and intronic splicing enhancers (58) are also reported. If and how these phenomena are correlated to RNA structural scaffold will require further investigation. It will also be worth investigating if structurally related proteins of both U2AF65 (PUF60, RBM39, and RBM23) and U2AF35 (U2AF26, ZRSR2P1, and ZRSR2) (59) could replace U2AF65 and U2AF35 and provide a yet another layer of complexity to splicing regulation. Moreover, the ability of U1 snRNP to recognize a structural landscape beyond 5′SS-like sequence could be explored for understanding its binding mechanism to non-coding RNAs (60).

Overall, the results described above suggest that splicing motifs could be structurally and functionally integrated within a splicing-conducive three-dimensional pre-mRNA structural scaffold. Therefore, it may not be possible to comprehensively assign splicing code to individual sequences without considering their presentation within the pre-mRNA scaffold. This work provides the initial

base to understand the contribution of pre-mRNA global structural framework to the mammalian splicing code.

REFERENCES

1. Will, C.L. and Luhrmann, R. (2011) Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.*, **3**, pii: a003707.
2. Ule, J. and Blencowe, B.J. (2019) Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. *Mol Cell*, **76**, 329-345.
3. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B. *et al.* (2019) Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, **176**, 535-548.e524.
4. Mayerle, M., Yitiz, S., Soulette, C., Rogel, L.E., Ramirez, A., Ragle, J.M., Katzman, S., Guthrie, C. and Zahler, A.M. (2019) Prp8 impacts cryptic but not alternative splicing frequency. *Proc Natl Acad Sci U S A*, **116**, 2193-2199.
5. Semlow, D.R., Blanco, M.R., Walter, N.G. and Staley, J.P. (2016) Spliceosomal DEAH-Box ATPases Remodel Pre-mRNA to Activate Alternative Splice Sites. *Cell*, **164**, 985-998.
6. Baralle, M. and Baralle, F.E. (2018) The splicing code. *Bio Systems*, **164**, 39-48.
7. McManus, C.J. and Graveley, B.R. (2011) RNA structure and the mechanisms of alternative splicing. *Curr Opin Genet Dev*, **21**, 373-379.
8. Fu, X.D. and Ares, M., Jr. (2014) Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet*, **15**, 689-701.
9. Baeza-Centurion, P., Miñana, B., Schmiedel, J.M., Valcárcel, J. and Lehner, B. (2019) Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell*, **176**, 549-563.e523.
10. Hiller, M., Zhang, Z., Backofen, R. and Stamm, S. (2007) Pre-mRNA Secondary Structures Influence Exon Recognition. *PLOS Genetics*, **3**, e204.
11. Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J. and Frey, B.J. (2010) Deciphering the splicing code. *Nature*, **465**, 53-59.
12. Corley, M., Burns, M.C. and Yeo, G.W. (2020) How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Mol Cell*, **78**, 9-29.
13. Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Van Nostrand, E.L., Pratt, G.A. *et al.* (2018) Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell*, **70**, 854-867.e859.
14. Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G.W., Jr, M.A. and Fu, X.-D. (2013) Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol. Cell*, **50**, 223-235.
15. Ohno, K., Takeda, J.I. and Masuda, A. (2018) Rules and tools to predict the splicing effects of exonic and intronic mutations. *Wiley Interdiscip Rev RNA*, **9**.
16. Lee, Y. and Rio, D.C. (2015) Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu Rev Biochem*, **84**, 291-323.
17. Saha, K., England, W., Fernandez, M.M., Biswas, T., Spitale, R.C. and Ghosh, G. (2020) Structural disruption of exonic stem–loops immediately upstream of the intron regulates mammalian splicing *Nucleic Acids Res*, **48**, 6294-6309.
18. Wang, J., Smith, P.J., Krainer, A.R. and Zhang, M.Q. (2005) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res*, **33**, 5053-5062.
19. Sanford, J.R., Wang, X., Mort, M., VanDuyn, N., Cooper, D.N., Mooney, S.D., Edenberg, H.J. and Liu, Y. (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res*, **19**, 381-394.
20. Änkö, M.L., Müller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J. and Neugebauer, K.M. (2012) The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol*, **13**, R17.

21. Zhou, Z. and Fu, X.-D. (2013) Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma*, **122**, 191-207.
22. Jobbins, A.M., Reichenbach, L.F., Lucas, C.M., Hudson, A.J., Burley, G.A. and Eperon, I.C. (2018) The mechanisms of a mammalian splicing enhancer. *Nucleic Acids Res*, **46**, 2145-2158.
23. Wong, M.S., Kinney, J.B. and Krainer, A.R. (2018) Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites. *Mol Cell*, **71**, 1012-1026.e1013.
24. Li, X., Liu, S., Zhang, L., Issaian, A., Hill, R.C., Espinosa, S., Shi, S., Cui, Y., Kappel, K., Das, R. *et al.* (2019) A unified mechanism for intron and exon definition and back-splicing. *Nature*, **573**, 375-380.
25. Abovich, N. and Rosbash, M. (1997) Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell*, **89**, 403-412.
26. Bennett, M., Michaud, S., Kingston, J. and Reed, R. (1992) Protein components specifically associated with prespliceosome and spliceosome complexes. *Genes Dev*, **6**, 1986-2000.
27. Zuo, P. and Maniatis, T. (1996) The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes Dev*, **10**, 1356-1368.
28. Kondo, Y., Oubridge, C., Roon, A.M.v. and Nagai, K. (2015) Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife*, **4**, DOI: 10.7554/eLife.04986.
29. Cho, S., Hoang, A., Sinha, R., Zhong, X.-Y., Fu, X.-D., Krainer, A.R. and Ghosh, G. (2011) Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly. *Proc. Natl. Acad. Sci. USA*, **108**, 8233-8238.
30. Zhou, Z., Sim, J., Griffith, J. and Reed, R. (2002) Purification and electron microscopic visualization of functional human spliceosomes. *Proc Natl Acad Sci U S A*, **99**, 12203-12207.
31. Michaud, S. and Reed, R. (1993) A functional association between the 5' and 3' splice site is established in the earliest prespliceosome complex (E) in mammals. *Genes Dev.*, **7**, 1008-1020.
32. Cho, S., Hoang, A., Chakrabarti, S., Huynh, N., Huang, D.B. and Ghosh, G. (2011) The SRSF1 linker induces semi-conservative ESE binding by cooperating with the RRMs. *Nucleic Acids Res*, **39**, 9413-9421.
33. Smola, M.J., Rice, G.M., Busan, S., Siegfried, N.A. and Weeks, K.M. (2015) Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nature protocols*, **10**, 1643-1669.
34. Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L. and Dreyfuss, G. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, **468**, 664-668.
35. Zhou, Z. and Reed, R. (2003) Purification of functional RNA-protein complexes using MS2-MBP. *Current protocols in molecular biology*, **Chapter 27**, Unit 27.23.
36. Mathews, D.H. (2014) RNA secondary structure analysis using RNAstructure. *Curr. Protoc. Bioinformatics*, **46**, 12.16.11-12.16.25.
37. Kohtz, J.D., Jamison, S.F., Will, C.L., Zuo, P., Lührmann, R., Garcia-Blanco, M.A. and Manley, J.L. (1994) Protein-protein interactions and 5′-splice-site recognition in mammalian mRNA precursors. *Nature*, **368**, 119-124.
38. Krainer, A.R., Maniatis, T., Ruskin, B. and Green, M.R. (1984) Normal and mutant human β-globin pre-mRNAs are faithfully and efficiently spliced *in vitro*. *Cell*, **36**, 993-1005.
39. Pineda, J.M.B. and Bradley, R.K. (2018) Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev*, **32**, 577-591.
40. Tarn, W.Y. and Steitz, J.A. (1994) SR proteins can compensate for the loss of U1 snRNP functions in vitro. *Genes Dev*, **8**, 2704-2717.
41. Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q. and Krainer, A.R. (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*, **31**, 3568-3571.

42. Wang, W., Maucuer, A., Gupta, A., Manceau, V., Thickman, K.R., Bauer, W.J., Kennedy, S.D., Wedekind, J.E., Green, M.R. and Kielkopf, C.L. (2013) Structure of phosphorylated SF1 bound to U2AF(6)(5) in an essential splicing factor complex. *Structure (London, England : 1993)*, **21**, 197-208.
43. Guth, S. and Valcárcel, J. (2000) Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *J Biol Chem*, **275**, 38059-38066.
44. Crawford, D.J., Hoskins, A.A., Friedman, L.J., Gelles, J. and Moore, M.J. (2013) Single-molecule colocalization FRET evidence that spliceosome activation precedes stable approach of 5' splice site and branch site. *Proc Natl Acad Sci U S A*, **110**, 6783-6788.
45. Kim, S.W., Taggart, A.J., Heintzelman, C., Cygan, K.J., Hull, C.G., Wang, J., Shrestha, B. and Fairbrother, W.G. (2017) Widespread intra-dependencies in the removal of introns from human transcripts. *Nucleic Acids Res*, **45**, 9503-9513.
46. Girard, C., Will, C.L., Peng, J., Makarov, E.M., Kastner, B., Lemm, I., Urlaub, H., Hartmuth, K. and Lührmann, R. (2012) Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nat Commun*, **3**, 994.
47. Taliaferro, J.M., Lambert, N.J., Sudmant, P.H., Dominguez, D., Merkin, J.J., Alexis, M.S., Bazile, C. and Burge, C.B. (2016) RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. *Mol Cell*, **64**, 294-306.
48. Hollander, D., Naftelberg, S., Lev-Maor, G., Kornblihtt, A.R. and Ast, G. (2016) How Are Short Exons Flanked by Long Introns Defined and Committed to Splicing? *Trends in genetics : TIG*, **32**, 596-606.
49. Zhang, S., Aibara, S., Vos, S.M., Agafonov, D.E., Lührmann, R. and Cramer, P. (2021) Structure of a transcribing RNA polymerase II-U1 snRNP complex. *Science*, **371**, 305-309.
50. Martinez-Contreras, R., Fisette, J.F., Nasim, F.U., Madden, R., Cordeau, M. and Chabot, B. (2006) Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol*, **4**, e21.
51. Liao, S.E. and Regev, O. (2020) Splicing at the phase-separated nuclear speckle interface: a model. *Nucleic Acids Res*.
52. Sharma, S., Kohlstaedt, L.A., Damianov, A., Rio, D.C. and Black, D.L. (2008) Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat Struct Mol Biol*, **15**, 183-191.
53. Hodson, M.J., Hudson, A.J., Cherny, D. and Eperon, I.C. (2012) The transition in spliceosome assembly from complex E to complex A purges surplus U1 snRNPs from alternative splice sites. *Nucleic Acids Res*, **40**, 6850-6862.
54. Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., Zhou, J., Qiu, J., Jiang, L., L, i.H. *et al.* (2014) Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. *Nat Struct Mol Biol*, **21**, 997-1005.
55. Soares, L.M., Zanier, K., Mackereth, C., Sattler, M. and Valcárcel, J. (2006) Intron removal requires proofreading of U2AF/3' splice site recognition by DEK. *Science*, **312**, 1961-1965.
56. Tavanez, J.P., Madl, T., Kooshapur, H., Sattler, M. and Valcárcel, J. (2012) hnRNP A1 proofreads 3' splice site recognition by U2AF. *Mol Cell*, **45**, 314-329.
57. Page-McCaw, P.S., Amonlirdviman, K. and Sharp, P.A. (1999) PUF60: a novel U2AF65-related splicing activity. *Rna*, **5**, 1548-1560.
58. Wei, W.J., Mu, S.R., Heiner, M., Fu, X., Cao, L.J., Gong, X.F., Bindereif, A. and Hui, J. (2012) YB-1 binds to CAUC motifs and stimulates exon inclusion by enhancing the recruitment of U2AF to weak polypyrimidine tracts. *Nucleic Acids Res*, **40**, 8622-8636.
59. Wu, T. and Fu, X.D. (2015) Genomic functions of U2AF in constitutive and regulated splicing. *RNA Biol*, **12**, 479-485.

60. Engreitz, J.M., Sirokman, K., McDonel, P., Shishkin, A.A., Surka, C., Russell, P., Grossman, S.R., Chow, A.Y., Guttman, M. and Lander, E.S. (2014) RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell*, **159**, 188-199.

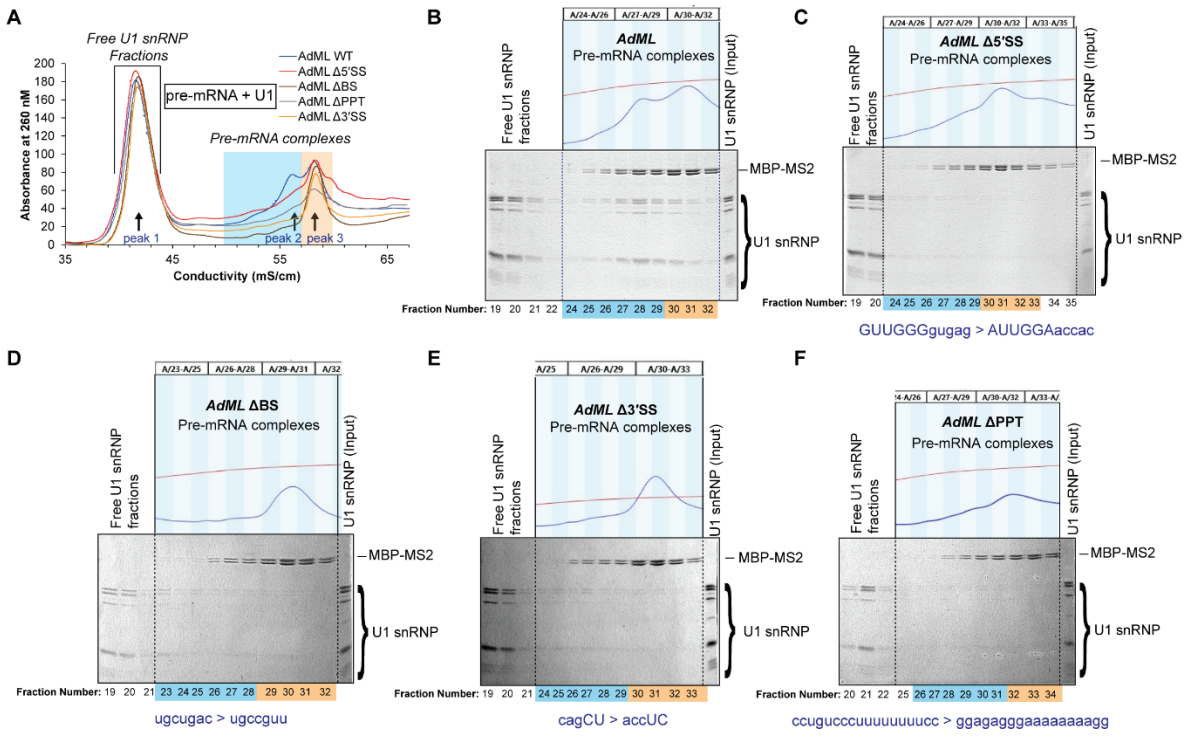


Figure 1. *AdML* splice signal mutants with intact 5′SS are defective in U1 snRNP recruitment

(A) Chromatograms of purification of U1 snRNP:pre-mRNA:3XMBP-MS2 complexes by anion-exchange chromatography with an NaCl gradient; conductivity instead of volume is plotted along *x*-axis to reduce shifts in peak positions caused by inconsistency in salt gradient generation by FPLC; peaks 1, 2, 3 of WT *AdML* chromatogram (a clear peak 2 is absent in the chromatograms of *AdML* mutants) are indicated with arrows; each of the fractions within the region indicated to contain 'pre-mRNA complexes' were concentrated by pull-down with amylose resin for analysis by Coomassie-stained SDS PAGE; blue shade corresponds to peak 2 and preceding fractions of *AdML* WT complexes and equivalent fractions of *AdML* mutant complexes; orange shade corresponds to fractions within peak 3. (B, C, D, E, F) SDS PAGE of *AdML* WT (B), *AdML* Δ5′SS (C), *AdML* ΔBS (D), *AdML* Δ3′SS (E), and *AdML* ΔPPT (F) fractions corresponding to the chromatogram shown in (A); 2.5 pmol U1 snRNP was used in the 'input' lane in the SDS gel; blue

and orange shade shown at the bottom of the gel images corresponds to fractions within the region of the chromatogram shaded with the same color; fraction-numbers of the analyzed samples are provided under each lane; mutated sequence for *AdML* mutants is also indicated under the SDS gel images; the raw chromatogram representing the fractions containing the pre-mRNA complexes are shown above each gel, where the red line indicates the conductivity trace.

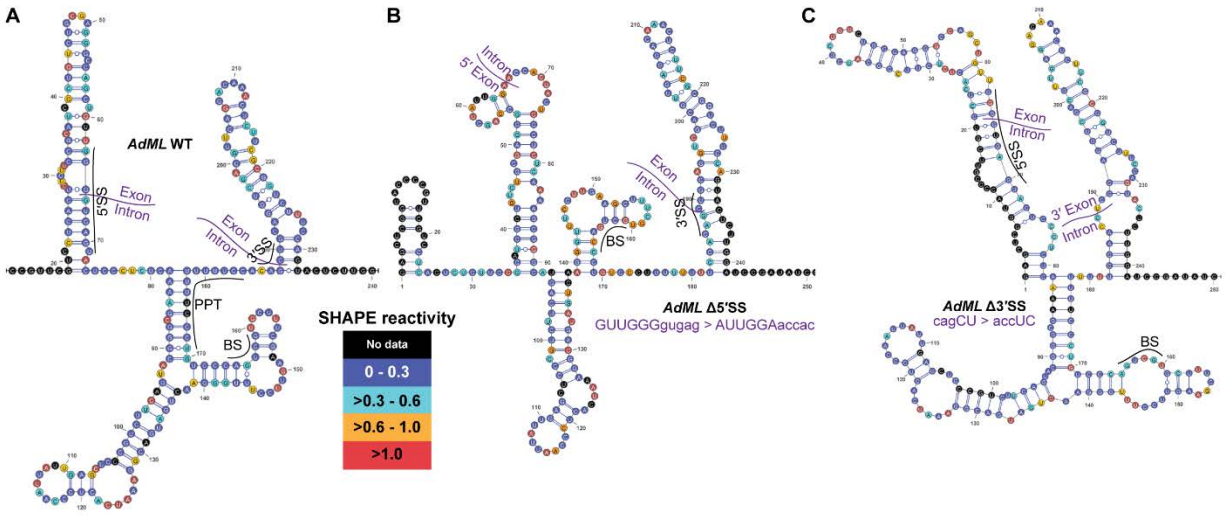


Figure 2. Mutation in major splice signals disrupts global structure of *AdML*

(A, B, & C) SHAPE-derived secondary structure models of *AdML* WT (A), *AdML* Δ5′SS (B), and *AdML* Δ3′SS (C); nucleotides are color-coded according to their SHAPE reactivity (see associated legend); positions of splice signals are indicated; position of PPT is shown only with the WT substrate for clarity; the mutated sequence is shown with each model.

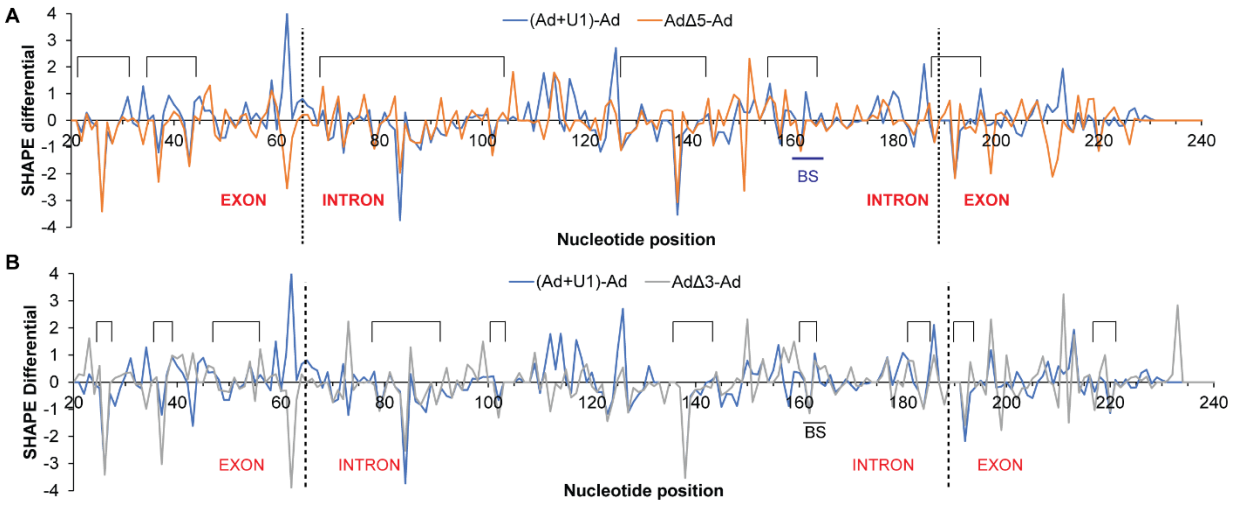


Figure 3. Strandedness of nucleotides across the pre-mRNA correlates to U1 snRNP recruitment

(A) Differentials of SHAPE reactivity of WT *AdML*+U1 snRNP complex and protein-free *AdML* i.e. [(Ad+U1) – Ad] and *AdML* Δ5′SS and *AdML* WT i.e. [Ad Δ5 – Ad] are overlaid onto each other; nucleotide positions exhibiting distinct negative values in both curves are indicated with rectangular bracket above the plot (these nucleotides loose SHAPE reactivity upon engagement of U1 snRNP potentially due to contact with U1 snRNP in *AdML* WT and are occluded by base-pairing in the protein-free *AdML* Δ5′SS). (B) SHAPE reactivity differentials of WT *AdML*+U1 snRNP complex and protein-free *AdML* i.e. [(Ad+U1) – Ad] and *AdML* Δ3′SS and *AdML* WT i.e. [Ad Δ3 – Ad] are overlaid onto each other.

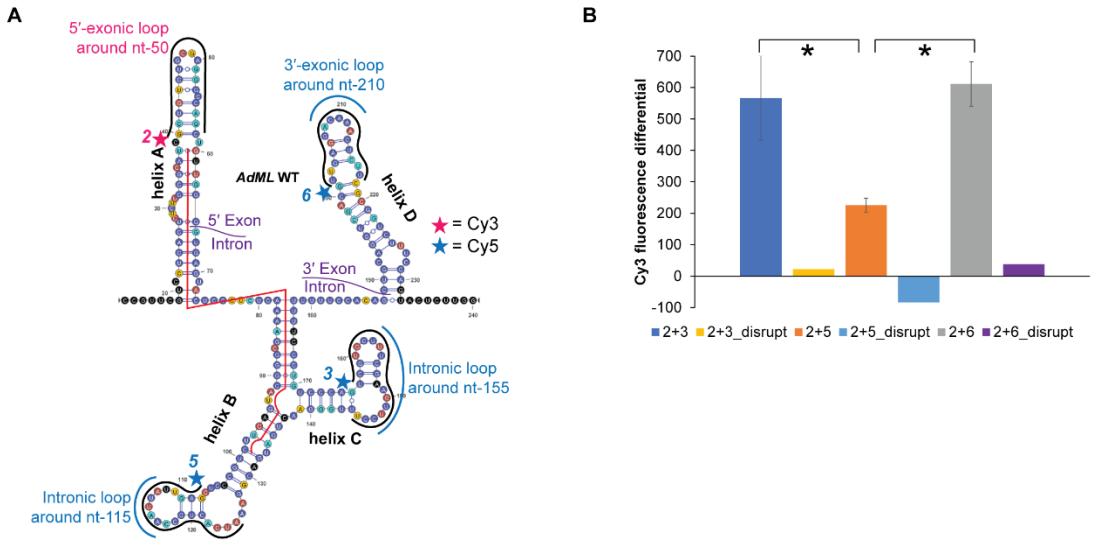


Figure 4. Close spatial positioning of splice signals to 5′SS

(A) SHAPE-derived secondary structure models of WT *AdML* showing the positions of the Cy3- and Cy5-labeled DNA probes; Cy3-labeled DNA probe is labeled as 2 and the remaining probes as 5, 3, and 6; the position of the 40-nt long structure-disrupting DNA is shown as a red line; the four helices of *AdML* are labeled as A, B, C, D. (B) Reduction in Cy3 emission upon hybridization of Cy5 probe to protein-free *AdML* WT shown for probe-pairs 2-3, 2-5, and 2-6; error bars indicate standard deviation (n=3); '*' indicates statistical significance ($p < 0.05$) obtained by one-tailed t-test; the same differential is also shown for *AdML* WT annealed to the disruptor DNA (n=1).

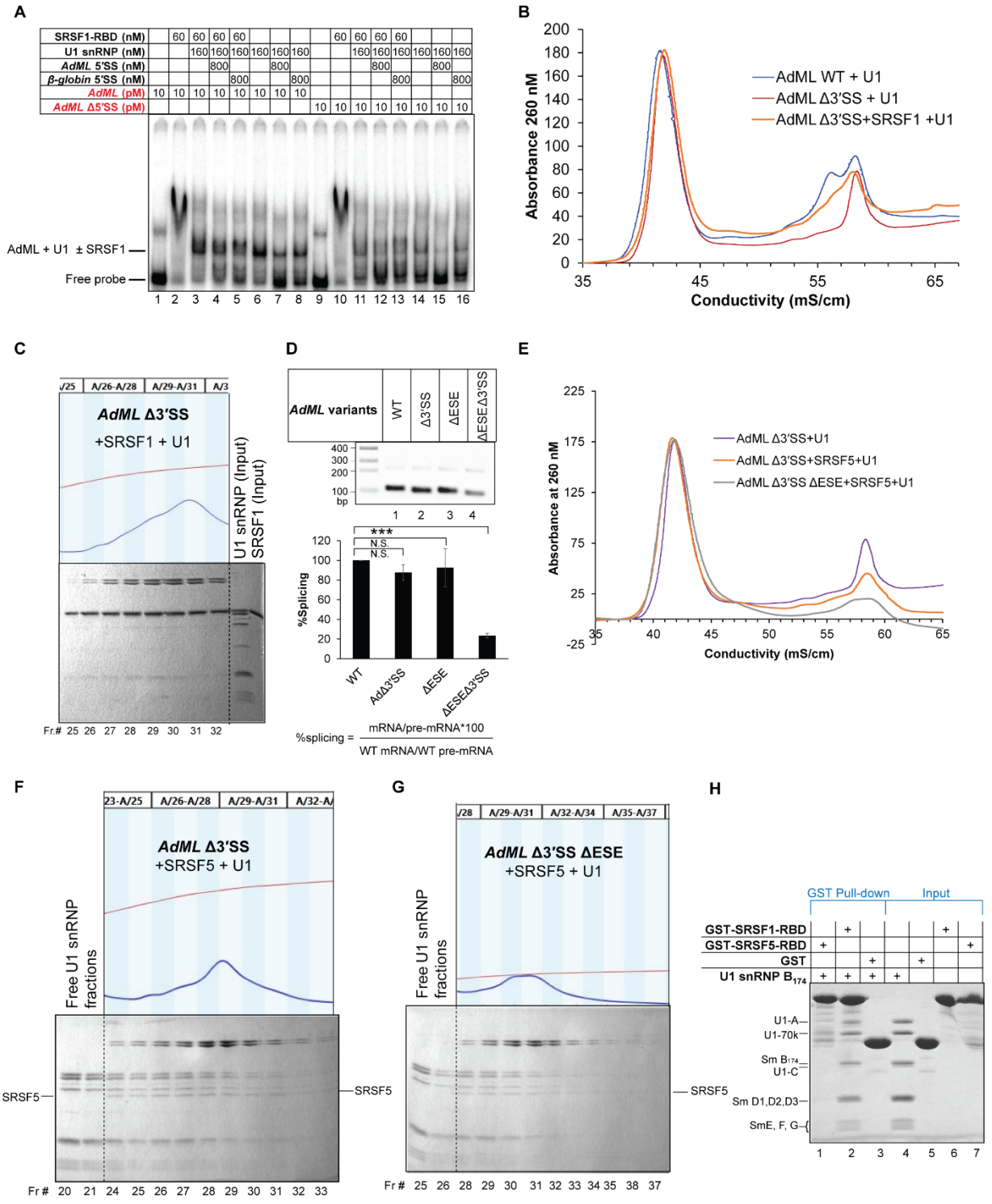


Figure 5. Differential functionalities of SR proteins with *AdML* WT and *AdML* Δ3′SS

(A) Electrophoretic mobility shift assay comparing the stability of U1 snRNP:*AdML* complex formed in the presence (lanes 3, 4, 5) or the absence (lanes 6, 7, 8) of SRSF1 (an ESE-dependent RBP reported to stabilize U1 snRNP on the pre-mRNA by directly interacting with U1 snRNP) by challenging the complexes with short 14-nt 5′SS RNA to visualize the level of released free radiolabeled probe; all U1 snRNP-dependent complexes formed more weakly with *AdML* Δ5′SS (lanes 9-16); red script indicates radioactive components in the gel. (B) Chromatogram of purification of U1 snRNP-dependent complexes formed with *AdML* Δ3′SS in the presence of SRSF1; the chromatograms of purification U1 snRNP-dependent complexes formed with protein-free *AdML* WT and protein-free *AdML* Δ3′SS are shown for comparison. (C) SDS PAGE analysis of the pre-mRNA complexes formed with U1 snRNP and *AdML* Δ3′SS in the presence of SRSF1; full-length phosphomimetic SRSF1 stains five times more intensely than U1 snRNP proteins U1-A and U1-70k (Supplementary Figure S6I). (D) (Top) Transfection-based splicing assay of *AdML* WT and Δ3′SS with or without ESE mutation (mutated region shown in Supplementary Figure S5E); (bottom) %splicing calculated from densitometric estimation of mRNA and pre-mRNA bands using the indicated formula of three biological replicas; error bar indicates standard deviation; N.S. = not significant, *** = $p < 0.005$ (one tailed one sample t-test). (E) Chromatograms of U1 snRNP-dependent complexes formed with *AdML* Δ3′SS and *AdML* Δ3′SS ΔESE in the presence of SRSF5; chromatogram of purification of mixture of U1 snRNP and protein-free *AdML* Δ3′SS is shown for comparison. (F, G) SDS PAGE of fractions corresponding to the pre-mRNA complexes formed with *AdML* Δ3′SS in the presence of U1 snRNP and SRSF5 (F), with *AdML* Δ3′SS ΔESE in the presence of U1 snRNP and SRSF5 (G); the raw SDS gel images are provided in Supplementary File 3. (H) GST-pull down assay showing absence of detectable interaction between U1 snRNP and GST-SRSF5 (lane 1) and the presence of interaction between U1 snRNP and GST-SRSF1-RBD (lane 2); Sm B_{174} indicates Sm B (1-174 a.a.).

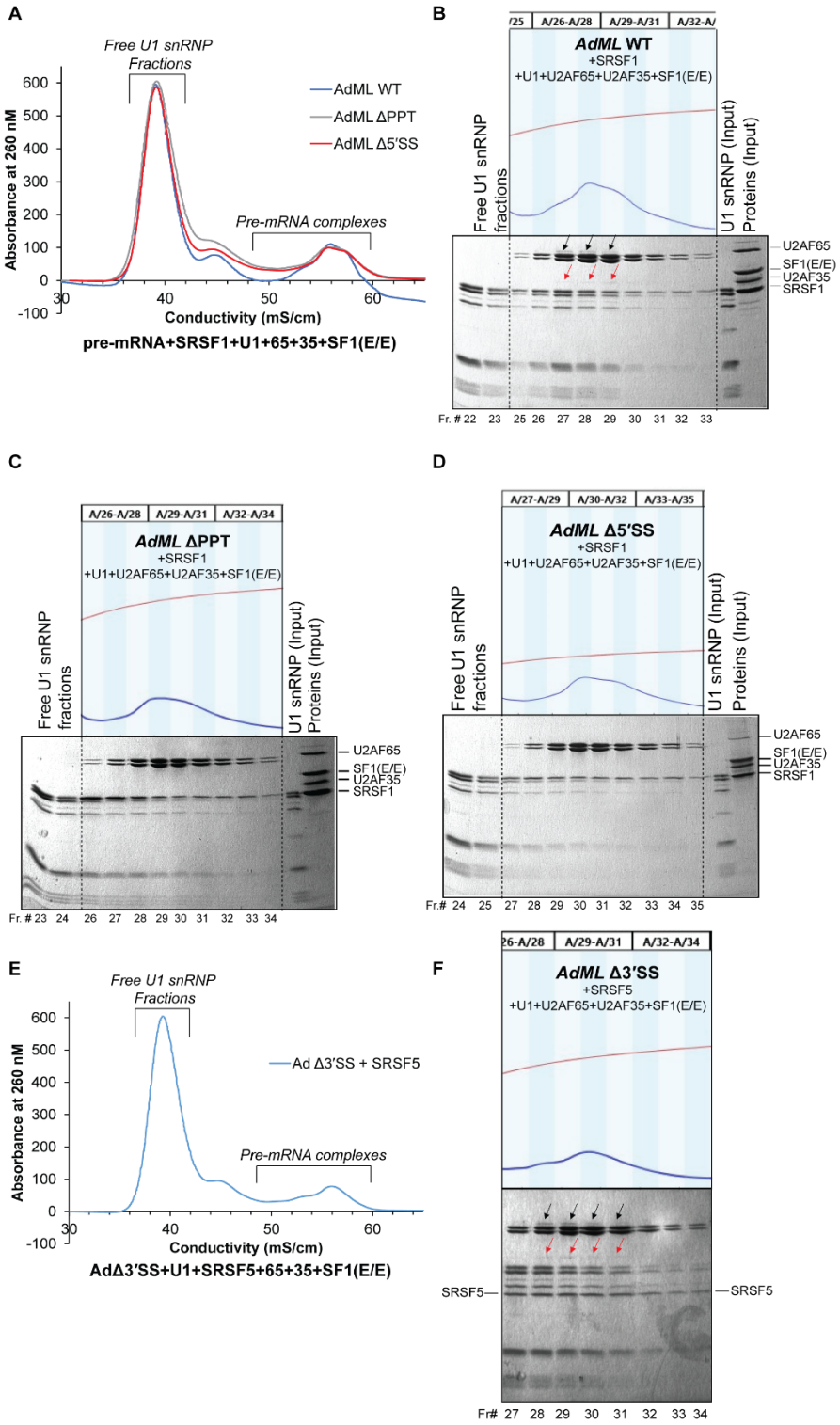
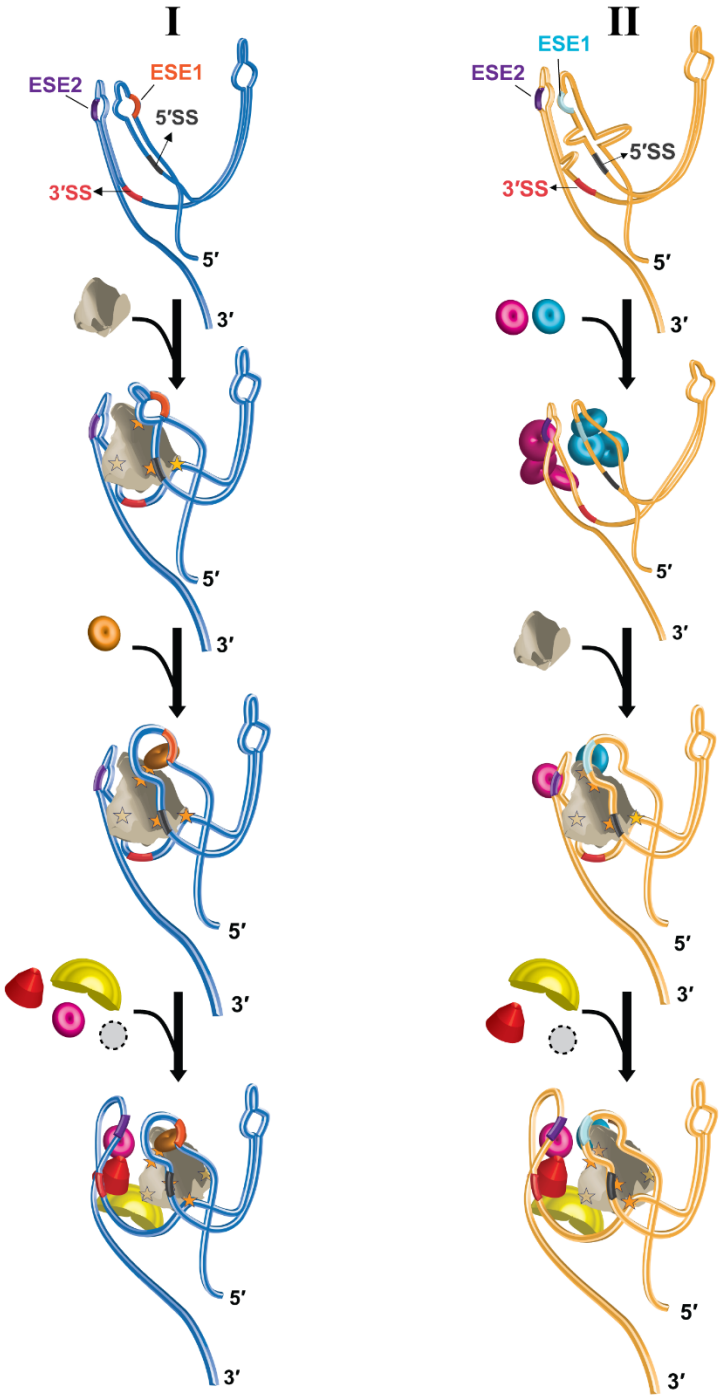


Figure 6. Interplay between ESE and global structure of *AdML* variants regulates recruitment of U2AF65/U2AF35

(A) Chromatograms of purification of *AdML* complexes formed with WT, Δ5′SS, and ΔPPT in the presence of U1 snRNP, SRSF1, U2AF65, U2AF35, and SF1 (S80E/S82E); the raw chromatograms including fraction numbers are shown above each gel. (B, C, D) SDS PAGE of fractions corresponding to chromatographically purified *AdML* WT complexes (B), ΔPPT complexes (C), Δ5′SS complexes (D); 'black arrow' indicates protein bands of U2AF65 and 'red arrow' U2AF35; the chromatography fractions marked as pre-mRNA complexes in A are formed with MBP-MS2-bound pre-mRNAs and concentrated by amylose pull-down before SDS PAGE; fraction numbers are given under each SDS gel; raw SDS gel images are provided in Supplementary File 3. (E) Chromatogram of purification of complexes formed with *AdML* Δ3′SS, U1 snRNP, SRSF5, U2AF65, U2AF35, and SF1 (S80E/S82E). (F) SDS PAGE of fractions containing the pre-mRNA complexes corresponding to the chromatogram shown in E; the positions of U2AF65 and U2AF35 bands are marked with arrows as in B; the position of SRSF5 band is indicated.



Pre-mRNA structure engages U1 snRNP and then stabilizes it in an ESE-dependent manner

Pre-mRNA structure is unfavorable for engaging U1 snRNP and requires to be structurally remodeled prior to U1 snRNP recruitment

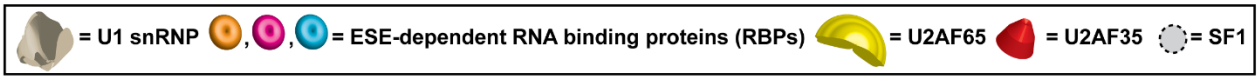


Figure 7. Proposed model of splicing substrate activation by cooperation between global pre-mRNA structure and ESEs

Pathway I: The global three-dimensional structure of a pre-mRNA enables engagement (strong or weak) of U1 snRNP; the splice sites and ESEs in the 5′ and 3′ exons are indicated; PPT and BS immediately upstream of the 3′SS are not drawn for clarity. Four hypothetical contact points between U1 snRNP and the protein-free pre-mRNA including the one at 5′SS region are shown with yellow stars; the solid stars suggest contact on the top surface and the translucent stars the opposite surface. Recruitment of an RBP at an ESE of the 5′ exon further stabilizes U1 snRNP on the pre-mRNA – the additional contact point between the RBP and U1 snRNP is shown with a star. This complex has the right conformation to receive U2AF65 and U2AF35, which are further stabilized by an RBP bound at an ESE in the 3′ exon. SF1 may play a chaperone-like role in the latter process. **Pathway II:** The pre-mRNA structure is unfavorable for engagement of U1 snRNP and needs to be structurally remodeled by additional factors such as ESE-dependent RBPs. The ESE-dependent RBPs could bind the pre-mRNA in multiple copies cooperatively, which structurally remodels the pre-mRNA (17), enabling recruitment of U1 snRNP (the contact points between U1 snRNP and the pre-mRNA and that between U1 snRNP and the ESE-bound RBP are indicated with five stars). This complex then may receive U2AF65 and U2AF35.