bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Exonic unpaired elements modulate pre-mRNA structure for splice site recognition

Kaushik Saha[1], Tapan Biswas[1], Mike Minh Fernandez[1], Whitney England[2], Robert Spitale[2], Gourisankar Ghosh[1,*]

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

[2]Department of Pharmaceutical Sciences, University of California Irvine, 147 Bison Modular, Building 515, Irvine, CA 92697

*Correspondence: gghosh@ucsd.edu

ABSTRACT

Splicing is a predominant process underlying the regulation of gene expression. The splice sites at the exon-intron boundaries are initially recognized by early spliceosomal components despite having a weak sequence conservation. It is unclear how the conserved splicing machinery distinguishes the authentic splice sites from a vast number of potentially similar but incorrect sites. Here we show that single-stranded RNA elements immediately upstream of the 5' splice site, regardless of nucleotide sequence, mediate functional interactions of pre-mRNAs with serine-arginine-rich proteins to modulate the pre-mRNA structure; this specifies the splice sites for recognition by the early spliceosome. This indicates that splice sites are defined not just by their primary sequence but also through single-stranded exonic elements within the protein-coding region of nascent pre-mRNAs.

INTRODUCTION

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

About a third of known genetic disorders are linked to splicing (1). Furthermore, the expression of quantitative trait loci for various complex diseases involve splicing (2). Several synonymous mutations causing a multitude of diseases including cancer are generally not mechanistically well-characterized (3). These facts overwhelmingly underscore the need to understand the mechanism of splice site recognition, but our current knowledge is rather limited.

The human spliceosome consists of five uridine-rich ribonuclear protein particles (U snRNPs) and several non-snRNP proteins (4). In addition to 5′ and 3′ splice sites (SS) at the ends of the intron, the splicing machinery recognizes two other major sequences within the intron: the branch-point site (BS) upstream of 3′SS, and a polypyrimidine tract (PPT) between BS and 3′SS. U1 snRNP, SF1, U2AF65, and U2AF35 recognize and bind the 5′SS, BS, PPT, and 3'SS, respectively, in the first step of spliceosome assembly leading to formation of the E-complex, the first known spliceosomal complex. However, limited sequence conservation observed in these splice signals implies that the recognition of these *cis*-acting sequences by *trans*-acting splicing factors gains external cues outside of their nucleotide sequences, possibly from structural elements embedded within the pre-mRNAs themselves.

The kinetic competition between splicing and transcription elongation (5, 6) has been proposed to regulate both constitutive and alternative splicing (7, 8). The rate of transcription elongation is considered to be a regulatory parameter for co-transcriptional RNA folding (9) and splicing (10). However, our knowledge of the effect of pre-mRNA folding on splicing is limited to case specific studies of a small number of pre-mRNAs (11-15). Several recent investigations indicate that transcripts form a variety of secondary and tertiary structures *in vivo* (16-20); transcripts are not immediately coated by proteins as previously thought and protein binding is dependent on the context and RNA structure (21, 22). These reports underscore the need and provide a groundwork for analyzing the role of pre-mRNA structure in constitutive splicing.

Serine-arginine-rich (SR) proteins have been implemented in both constitutive and alternative splicing (23). SR proteins belong to a family of RNA-binding proteins that contain an N-terminal RNA-

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

binding domain (RBD) with either one or two RNA recognition motifs (RRM), and an Arg-Ser-rich (RS) domain at the C-terminus. Co-transcriptional engagement of SR proteins to the nascent RNA polymerase II transcripts has been demonstrated *in vivo* with GFP-tagged SR proteins (24) but little information is available that offers insight into how SR protein binding to the pre-mRNAs is regulated or what are the consequences of this binding.

In this study, we show that single-stranded exonic segments immediately upstream of the 5′SS mediate functional interactions of SR proteins to nascent pre-mRNAs leading to structural modulation of the pre-mRNA, which is essential for initiation of spliceosome assembly. Thus, nucleotides within splice sites and exonic unpaired elements cooperate with each other and other splice signals for exon-intron definition and initiation of spliceosome assembly.

RESULTS

Single-stranded exonic elements immediately upstream of 5′SS promote splicing regardless of whether 5′SS is single-stranded

In order to gain understanding into whether and how pre-mRNA secondary structure influences constitutive splicing, we analyzed secondary structure models of two model pre-mRNA substrates, namely human *β-globin* IVS1 (25) and Adenovirus 2 major late transcription unit IVS1 (*AdML*) (26) by measuring the *in vitro* SHAPE reactivity in buffer (27) (Fig. 1A and Fig. S1A). SHAPE measures the 2'-hydroxyl flexibility and, as such, is a proxy for the identification of single-stranded nucleotides within folded RNAs. Both substrates were organized into three stem-loop constituents: a significant portion of the 5′ exon and 5′ end of the intron formed the first component, the rest of the intron formed the second portion, and the 3′ exon formed the third.

To determine the role of the exonic and intronic stem-loops in splicing, we carried out mutagenesis to hybridize one strand of the loops to the opposite strand and then examined the resulting splicing efficiency from the mutated templates. To examine whether the secondary structure adopted *in vitro*

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

had an impact on splicing *in vivo*, we used the transfection-based splicing assay in HeLa cells. Hybridization of loops in 5′ exons and introns into duplexes (hybridization-mutation) were named EH (exon hybridization) and IH (intron hybridization) mutants, respectively (Fig. 1A, 1B, Fig. S1A, S1B). All substrates were expressed under control of the CMV promoter for transcription *in vivo*. The transcription rate for different mammalian genes varies by several orders of magnitude with the median transcription rate being about 2 and the highest rate being about 500 mRNAs per hour. The transcription rate of CMV promoter is roughly 6-7 mRNAs per hour (28, 29). Therefore, the CMV promoter represents a strong promoter near the median strength of endogenous mammalian promoters.

In *β-globin*, hybridization of both terminal loops in the 5′SS-containing stem-loop (EH3+4) abolished splicing while hybridization of both loops opposite of 5′SS (EH1+2) caused ~30% loss of splicing (Fig. 1C, compare lanes 4 and 9). Because EH1+2, in which the 5′SS is completely hybridized, showed significantly greater efficiency in splicing, the defect observed in EH3+4 might not be caused by inaccessible 5′SS. Hybridization-mutations of the terminal loops in the 5′ exon of *AdML* also showed similar effects as in *β-globin* (Fig. S1C, lanes 1-3). *AdML*-EH1+2 produced a small quantity of aberrantly spliced RNA along with authentic mRNA; *AdML*-EH1+2+3+4 showed significant defects in splicing with the concomitant production of aberrantly spliced RNA. It is noteworthy that in *AdML*, the 5′SS is naturally hybridized. To examine if the hybridization of the terminal loops negatively impacts splicing in *AdML* by making 5′SS inaccessible, we hybridized the loops at the base of the 5′SS-containing the stem-loop (EH5, IH) (Fig. S1A). These loops display a much longer stretch of single-stranded region and hence hybridization-mutation of these loops is expected to increase the energy of unfolding of the 5′SS-containing stem-loops more efficiently than those of the much smaller terminal loops. Neither mutant showed any obvious defect in splicing (Fig. S1C, lanes 5-7).

To further understand the role of accessibility of 5′SS in splicing, we completely de-hybridized the 5′SS in splicing defective *β-globin* EH3+4 and *AdML* EH1+2+3+4 mutants and generated *β-globin* EH3+4+ED1+2 and *AdML* EH1+2+3+4+ED5 mutants (Fig. 1B, S1B). In both cases, the splicing defects

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

caused by hybridization of the terminal loops were not rescued through de-hybridization of the 5′SS (Fig. 1C, lanes 10-13, S1C, lanes 3-4). This strongly indicates that the single-stranded region immediately upstream of the 5'SS acts independent of base-pairing within the 5′SS.

Next, we examined if the complete absence of intermittent stems has any impact on the functionality of the exonic single-stranded element(s) immediately upstream of 5'SS. Lanes 14-16 of Fig. 1C show that ED1 or ED2 mutation in *β-globin* enhances splicing by about two-fold. This enhancement could be due to a more accessible 5'SS. Nonetheless, removal of base-pairing from the other stems (ED3, ED4) alone or in combination also did not impair splicing compared to the WT substrate (Fig. 1C, lanes 17-22). The reduction in splicing in ED1+2+3+4 from the elevated level of splicing of ED1 or ED2 could be due to loss of structural integrity in this region. ED3 and ED3+4 of *β-globin* displayed a small amount of an aberrant product (marked with an asterisk), indicating that base-pairing in the region immediately upstream of the 5'SS could act to prohibit splicing from cryptic splice sites. We removed base-pairing from the stems immediately upstream of the 5'SS in *AdML* (Fig. S1A) and observed that the transfection-based splicing assay did not show any significant defects (Fig. S1C, lanes 8-10).

We also deleted the loops 3+4 in *β-globin* producing *β-globin* ΔL mutant (Fig. 1A, 1B). Transfection-based splicing assay revealed an 80% loss of splicing efficiency (Fig. 1C, compare lanes 1 and 5) unlike EH3+4, which resulted in complete abolition of splicing (Fig. 1C, lane 4). Therefore, in contrast to EH3+4, ΔL did not completely abolish splicing. This is most likely because the ΔL deletion destabilizes the secondary structure masking the 5′SS by increasing free energy of folding of the entire stem-loop by ~ 7 kcal/mol and thus making the 5′SS more accessible. Making 5´SS more accessible has already been shown to enhance splicing of *β-globin* (ED1 or ED2 mutants; Fig. 1B, C, compare lanes 14, 15, 16). Therefore, the single-stranded regions within the 5' exon at or immediately upstream of 5'SS serve two purposes: first, they increase the accessibility of the 5'SS, and second, they promote splicing independent of hybridization observed in the 5'SS region.

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We finally compared the splicing activities of WT and EH3+4 constructs of *β-globin in vitro* with nuclear extract, which revealed sub-optimal splicing activity of the EH3+4 construct with concomitant production of several aberrant products (Fig. 1D, bands marked with * are aberrant products). Interestingly, detection of aberrant products was more conspicuous in the *in vitro* splicing assay than from the transfection-based splicing assay. This is likely due to a variety of reasons, one of which could be the inability of the primers used for RT-PCR to amplify the aberrant products in the transfection-based splicing assay. IH (intronic hybridization) mutants of *β-globin* (Fig. 1A) did not show any measurable defects in splicing (Fig. S1D).

We next tested splicing regulatory activity of the putative exonic loops of five additional constitutively spliced human pre-mRNAs of variable lengths: *SRSF1* IVS1, *SNRP70* IVS1, *ACTA2* IVS5, *HBA1* IVS1, and *SRSF7* IVS2. First, we inferred the secondary structure of their exon-intron junctions (encompassing -70-nt through +10-nt around 5′SS; see methods) and then converted the exonic loops into duplexes by mutagenesis (EH mutants). Fig. S2 (top panels) shows predicted secondary structure models of the target regions of the pre-mRNAs with the area of hybridization mapped onto it. The EH1 mutants of *SNRP70* IVS1, *ACTA2* IVS5, *HBA1* IVS1, and *SRSF7* IVS2 showed significant splicing defects (bottom panels of Fig. 3). 'RNAstructure' predicted two alternative structures of *SRSF1* IVS1 and based on these structures we generated two mutants EH1 and EH2. EH1 mutant generated based on the structure 1 showed an 80% loss of splicing while EH2 resulted in 25% loss of splicing (Fig. S2). A larger defect observed with the mutation based on structure 1 possibly indicates that structure 1 is the predominant structure of *SRSF1* exon 1 *in vivo*. The similar impacts of disrupting secondary structure within the pre-mRNA in the region immediately upstream of 5'SS on splicing both *in vitro* and *in vivo* strongly suggests that the nascent pre-mRNA assumes the shown structure in the target region *in vivo* and that this newly identified structural element immediately upstream of 5′SS guides the recruitment of essential splicing factors. We have termed the essential single-stranded segment(s)

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

immediately upstream of 5′SS the exonic unpaired element(s) or EUE, which could be either a continuous stretch of single-stranded RNA or a sequence with short intermittent stems.

The EUE displays no conservation or requirement in nucleotide sequence

In order to examine if EUE is a sequence-independent signal dictated by the number of available unpaired nucleotides, first we examined if the EUE mutants used for the splicing assays (such as *β-globin* EH3+4) have the expected secondary structure. Therefore, we determined the secondary structure model of *β-globin* EH3+4 mutant by *in vitro* SHAPE, which revealed base-pairing at expected positions (Fig. 2A). Next, we hybridized the opposite strands of all four loops of *β-globin* by mutating the 5′ strand of the stem-loop (EH1+2+3+4) (Fig. 1A, 1B, 2B, middle panel), which abolished splicing completely and produced an aberrant product (Fig. 2C, lane 2, aberrant product marked with an asterisk). Sanger sequencing of the aberrant product obtained from EH1+2+3+4 indicated that it has a 63-nt long region from within the 5' exon removed (shown between red borders in the middle panel of Fig. 2B). Then we introduced two loops opposite of 5′SS by mutating the 5′ strand of the stem (EH1+2+3+4+5L2) (Fig. 2B, left panel), which did not improve splicing significantly (Fig. 2C, lane 3). Next, we altered the sequence of the non-mutated strand of the terminal loop region to create a large terminal loop (EH1+2+3+4+TL) (Fig. 2B, right panel). This led to recovery of about 45% of splicing activity (Fig. 2C, lane 4). This result indicates that splicing capability does not require specific secondary structure in this region although splicing efficiency may depend on it. In all, these data indicate that the exonic region immediately upstream of 5′SS regulates constitutive splicing through a hitherto unexplained mechanism, which is not directly dependent on base-pairing within the 5′SS duplex, and that the single-stranded region immediately upstream of 5'SS does not require any specific nucleotide sequence to exert its effect. If the effects of available unpaired nucleotides on splicing efficiency is dependent on specific secondary structure of this region remains to be investigated in the future.

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To examine if the effects of the mutagenesis on splicing shown so far are correlated with alteration of nucleotide sequence, we substituted all purines with pyrimidines individually or in combinations within and around loops 3+4 (M1-M7) (Fig. 2D) without affecting the secondary structure. Interestingly, these mutants exhibited no discernible defects in splicing (Fig. 2C, lanes 5-7). Indeed, the secondary structures of the exonic stem-loops of WT *β-globin*, M1+2+3+4+6+7, and M1+2+3+4+6+7+8 mutants as predicted by 'RNAstructure' (30) were identical or almost identical to that of the SHAPE-derived secondary structure model of the exonic stem-loop of WT *β-globin* (Fig. S3A). An alignment of the substitution mutant M1+2+3+4+6+7+8 against the WT pre-mRNA reveals 66.6% mismatch between 70th and 100th nt (Fig. S3B). These results indicate that the single-stranded region immediately upstream of 5'SS promotes splicing regardless of its nucleotide sequence. In support of this observation, we generated a sequence logo from primary sequence alignment of 102 5' UTR exons spanning -70-nt through +10-nt around 5′SS from chromosome 1 (Fig. 2E). The complete absence of nucleotide sequence conservation within this wide range of exons negates any likelihood of a conserved nucleotide sequence in this region.

The exonic unpaired elements effectuate structural modulation of the nascent pre-mRNA *in vivo*

The results of the splicing assays with structural mutants of *β-globin* and *AdML* reported above (Fig. 1, 2, S1) indicate that the secondary structural features observed in protein-free pre-mRNA *in vitro* are also present in nascent pre-mRNA *in vivo*, at least within the region immediately upstream of the 5'SS. To verify this interpretation, we carried out *in vivo* DMS-footprinting of the 5′ exon-intron junction of transfected WT *β-globin*, and *β-globin* EH3+4. The scheme for this experiment is shown in Fig. 3A. For reverse transcription, primers were nested near the 5′ end of the intron to ensure that cDNA is synthesized from the pre-mRNA and not the mRNA. *In vivo*, reactivity of nucleotides was largely different from that observed *in vitro* with protein-free *β-globin*; high and low reactivity (indicative of high and low nucleotide flexibility) were observed in several stem and loop regions, respectively (compare

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the dot-bracket structural notation of the protein-free *β-globin* RNA variants with the *in vivo* DMS-reactivity; dot indicates single-stranded region and bracket base-paired region) (Fig. 3B). Surprisingly, the WT substrate showed significantly higher reactivity at various places than the mutant substrate within both exons and introns. Observation of *in vivo* flexibility of a nucleotide that is base-paired in the protein-free substrate indicates structural modulation. Enhanced flexibility of a nucleotide in the WT substrate compared to the EH3+4 mutant indicates that the EUE of *β-globin* is essential for structural modulation of the pre-mRNA *in vivo*.

We then carried out *in vivo* structural probing of an endogenous pre-mRNA, the endogenous *SRSF1* IVS1 synthesized from the genome of HeLa cells (not a transfected plasmid). The footprint again did not completely match with either structural model of *SRSF1* IVS1 predicted and tested by hybridization-mutation and splicing assay as shown in Fig. 3. Observation of nucleotide flexibility in regions where base-pairing is predicted in both structural models possibly indicates structural modulation (Fig. 3C, compare dot-bracket notation with the plot of DMS-reactivity). Nonetheless, high level of nucleotide flexibility immediately upstream of 5′SS is likely an indication of splicing competence of this pre-mRNA.

Exonic unpaired elements modulate the structure of the pre-mRNAs by mediating functional recruitment of SR proteins

SR proteins are the first among splicing factors to bind the pre-mRNA co-transcriptionally *in vivo* (24) and have been implemented in splice site recognition (23). Binding of specific SR proteins to protein-free pre-mRNAs has been shown to promote splicing *in vitro* (31). Several transcriptome-wide crosslinking and immunoprecipitation experiments followed by RNA-seq (CLIP-seq) identified that SR proteins bind the exonic region immediately upstream of 5'SS (32-34). eCLIP (enhanced CLIP) data (34) for binding of three different SR proteins to *SRSF1* exon 1 are shown in Fig. 4A (www.encodeproject.org). Therefore, we investigated the immediate consequence of SR protein binding to the model substrates and the bearing of the presence of EUE on the process. We tested the

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

effects of SRSF1 binding to *β-globin* and *AdML* and SRSF2 binding to *β-globin in vitro* by SHAPE reactivity. Both SR proteins are known to promote splicing of both substrates (31, 35, 36). Both gain and loss of SHAPE reactivity of individual nucleotides in both exons and introns were observed upon SR proteins binding in both *β-globin* (Fig. 4B, see top panel for SRSF1, bottom panel for SRSF2) and *AdML* (Fig. S4). Loss and gain of SHAPE reactivity of a nucleotide upon protein binding indicates loss and gain of nucleotide flexibility, respectively. Nucleotides would lose flexibility either due to bound protein molecules or by SR protein-induced base-pairing; they would gain flexibility due to removal of structural constraints such as base-pairing. Therefore, we conclude that SR protein binding induces modulation of the secondary structure in the pre-mRNA.

To understand if SR protein binding to the WT pre-mRNA differs from that of the splicing-defective EH3+4 (exon hybridization) mutants described before, we compared SHAPE reactivity of individual nucleotides of *β-globin* EH3+4 in the absence and presence of excess SRSF1 or SRSF2. We observed diminished gain of nucleotide flexibility in EH3+4 mutant upon binding of either SRSF1 or SRSF2 compared to that of the WT pre-mRNA (Fig. 4B). These results show that SR protein-mediated modulation of the pre-mRNA structure is dependent upon the presence of the EUE.

Structural modulation of the protein-free pre-mRNA by SR proteins led us to hypothesize that this structural modulation is essential for initial splice site recognition. To investigate this, we examined assembly of the E-complex, the first known spliceosomal complex within which both splice sites are recognized, with WT *β-globin* and its EH3+4 mutant by incubating the pre-mRNAs in nuclear extract depleted of ATP (37). We observed that the WT pre-mRNA formed unproductive H-complex instantly (at 0 min) and spliceosomal E-complex in 15 min (Fig. 4C, lanes 1-2). E-complex assembly was almost non-existent with the EH3+4 mutant (Fig. 4C, lanes 4-5). Binding of SR proteins to the exons and subsequent promotion of E-complex assembly have been documented before (32, 38). Our data show that the link between SR protein binding and E-complex assembly is SR protein-mediated EUE-dependent pre-mRNA structural modulation. In the accompanying manuscript, we have shown that SR

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

protein-mediated structural modulation leads to recognition of all splice signals of the model substrates by the respective splicing factors.

Exonic regions immediately upstream (~ 50-nt) of retained introns are highly structured genome wide

To correlate the secondary structure of the pre-mRNAs immediately upstream of 5′SS with splicing, we analyzed transcriptome-wide icSHAPE data (18) from mouse embryonic stem cells after reconstructing a splicing profile from these cells. The mRNAs that are analyzed are poly-A purified and hence, have already undergone splicing except where the introns are retained. The scheme of the *in vivo* and *in vitro* icSHAPE experiment is shown in Fig. 5A, 5B. Exons followed by retained introns displayed low SHAPE reactivity immediately upstream of the 5′SS both *in vitro* and *in vivo* (Fig. 5C, 5D). The low icSHAPE reactivity within the exons upstream of the retained introns *in vitro* and *in vivo* indicates that nucleotides in this region are mostly base-paired on average. In contrast, exons followed by the succeeding exon (i.e., with a spliced-out intron) showed high SHAPE reactivity *in vivo* and *in vitro* upstream of the 5′SS indicating that these RNAs are more flexible and likely have broad regions of single-stranded nucleotides. Interestingly, icSHAPE reactivity in the retained intron substrates is lowest between 15-nt and 50-nt immediately upstream of the 5′SS and not at the 5′SS, which indicates the said region to be mostly base-paired on average. An earlier report in which *in vivo* DMS-probing was employed to monitor splicing in *Arabidopsis thaliana* reported similar structural profiles of exons followed by retained introns (17). Therefore, it appears that complimentary structure probing methods reveal similar structural characteristics. The structural footprint of the exons followed by the retained introns are similar to the *in vivo* structural footprint of the transfected splicing-defective *β-globin* EH3+4 substrate, where the EUE is base-paired (Fig. 3B), and is in contrast to that of the transfected WT *β-globin* and endogenous *SRSF1* IVS1, where nucleotides immediately upstream of 5′SS are greatly flexible (Fig. 3B, 3C). Hence, these data correlate extensive base-pairing within the exonic region immediately upstream of 5′SS with

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

loss of splicing. However, the genome-wide data do not report whether the exons followed by spliced out introns had single-stranded segments immediately upstream of 5'SS before splicing (because the structural information was obtained after completion of splicing). Since our results with model substrates show that presence of the exonic unpaired elements immediately upstream of 5′SS is essential for pre-mRNA structural modulation and splicing, we hypothesize that splicing competent pre-mRNAs acquire the required single-stranded elements immediately upstream of the 5′SS either co-transcriptionally, through the action of a molecular switch as described before (39), or other as yet uncharacterized mechanisms. In addition, a previous genome-wide bioinformatic analysis of about 350,000 splice site regions indicates that the probability of intramolecular base-pairing within the pre-mRNA peaks at both 5'SS and 3′SS (40); this suggests that accessibility to the splice sites is likely dependent on pre-mRNA structural modulation in a wide range of pre-mRNAs.

We next compared the energy of folding of the exonic unpaired elements of the retained intron substrates with those of one-thousand 5′ UTR (untranslated region) exons. We first predicted secondary structure models of two different segments of 1000 UTR exons (spanning -70-nt through +10-nt, i.e. the target region, and -151-nt through -71-nt, i.e. the non-target region) from ten chromosomes (chr 1-10) using 'RNAstructure' (30). The median free energy of the target region was -17.2 kcal/mol while that for the non-target region was -19.5 kcal/mol (Fig. 5E). A linear regression analysis revealed a *p*-value for these energy differences to be 0.0047. We then estimated the energy of folding for the same region of the 291 retained intron substrates shown in Fig. 5D. The median free energy of folding for these exons is -22.3 kcal/mol (Fig. 5E). We surmise that the target region with the highest median free energy of folding, besides suggesting that it contains a higher proportion of unpaired nucleotides, is also indicative of a region with inherent flexibility and, therefore, a more likely candidate for initiating pre-mRNA structural modulation. The low free energy of folding (i.e. greater stability) in the target region of the 291 retained intron substrates is consistent with it having more stable secondary structure that could prevent recruitment of splicing factors and subsequent pre-mRNA structural modulation.

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCUSSION

The high sequence degeneracy observed in the *cis*-acting splice signals have long been a puzzle regarding how metazoan splice signals are initially recognized by splicing factors within early spliceosomal complexes, such as the E-complex. To explain how the splicing machinery recognizes only the authentic splice signals irrespective of the level of nucleotide conservation, it was proposed that splice signals are defined not only by nucleotide sequence but also by its 'context' (41). Our present work provides various key findings that clarifies this recognition mechanism and show that a structural context renders the splice signal motifs recognizable to the splicing factors. First, a segment immediately upstream of the 5'SS with single-stranded regions (either continuous or interrupted by short stems) is essential for splicing. Second, this region (referred to as EUE) mediates functional interactions between pre-mRNAs and serine-arginine-rich (SR) proteins, which modulate the pre-mRNA structure to initiate assembly of the early spliceosome. Third, the presence of simply an unpaired 5'SS is not sufficient for splicing, an EUE immediately upstream of the 5'SS is necessary. Fourth, we observed that restricting the potential for structural modulation of the pre-mRNA, such as within a highly structured EUE, renders a pre-mRNA incapable of splicing. Fifth, since binding of SR proteins modulates the structure of a naked pre-mRNAs *in vitro*, we surmise that this structural modulation is essential to 'specify' the splice signals in the nascent pre-mRNAs for recognition by an early spliceosomal complex, likely prior to the E-complex assembly. Indeed, in the accompanying manuscript, we have shown that the exonic unpaired elements are essential for recognition of all four major splice signals, leading up to the assembly of an early stable spliceosome, termed the recognition (R) complex, which precedes the assembly of the E-complex. Essentiality of structural modulation of the pre-mRNA likely implies that the splice sites could normally be inaccessible, perhaps through base-pairing as predicted through bioinformatic analyses of about 350,000 splice sites (40). Because in the model substrates, EUE is essential for splicing and assembly of the early spliceosomal complexes such as E-complex, a complex

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

considered universal for major spliceosome, we surmise that splicing competence of a wide range of pre-mRNAs depends on the presence of EUE. We hypothesize that pre-mRNAs could acquire EUE either co-transcriptionally, through the action of a molecular switch (39), or other hitherto unexplained mechanisms. Further investigations are required to examine how splicing competent pre-mRNAs obtain the exonic unpaired elements immediately upstream of 5′SS.

Our results identify a structural context of the splice sites essential for their recognition. These results also open many additional questions about the underlying mechanisms and players involved. We discovered that EUE upstream of 5′SS is essential for defining the context of constitutive splice sites whose nucleotide sequence resembles the consensus sequence and is considered strong indicating interdependency of the context and the nucleotide sequence for splice site recognition. Now the question remains how the functionality of EUE of the constitutively spliced pre-mRNAs remain uninterrupted regardless of biological variations. We propose that for an EUE to remain constitutively active, it must be activated by multiple SR proteins so that one SR protein can be compensated for by another in case of changes in SR protein expression level. Indeed, splicing of *β-globin*, a constitutively spliced substrate, can be promoted by one of three SR proteins (31, 42). There are known to be ten SR proteins available in humans. Therefore, it is not clear if promotion of splicing by only three SR proteins is sufficient for maintenance of constitutive splicing *in vivo*. Further investigations are required to clearly understand the mechanism of *in vivo* maintenance of constitutive splicing. On the other hand, it is possible that this EUE-mediated splice site recognition mechanism could play even a greater role for regulated (alternative) splicing. Since EUE could modulate splicing by regulating presentation of an RNA structural motif, it is likely to contribute to gene regulation under natural circumstances. We postulate that an EUE of a pre-mRNA undergoing regulated splicing gets activated exclusively in the presence of one or more splicing factors, such as a specific SR protein, and thus promote splicing only when that protein is expressed. In addition, since RNA structure is known to be altered by transcription elongation rate, temperature, activities of RNA chaperones or helicases, or other cellular processes, it

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is possible that alteration of EUE structure of specific pre-mRNAs is ensued under certain circumstances. A structural alteration causing increase or decrease in unpaired nucleotides availability in the EUE could alter its specificity for specific SR proteins or effectiveness in promoting splicing. Whether the specific nature of the secondary structure of EUE in addition to the number of available unpaired nucleotides is significant for splicing efficiency remains to be investigated.

The nature and role(s) of SR protein binding in context-dependent recruitment of splicing factors are not clear. CLIP-seq experiments indicate that the binding occurs more to regions immediately upstream of 5'SS and downstream of 3'SS, regardless of the nucleotide sequence of exons (32, 34). Our results tally to these results signifying importance of EUE, which is located immediately upstream of 5′SS, in SR-protein binding and consequent pre-mRNA structural modulation. However, we have little clue to the likely links between EUE and long-studied ESE (exonic splicing enhancer) motifs, cognate to SR proteins. Functional ESEs for specific SR proteins (43-45) and their positional preferences have been studied in great detail. Little correlation has been observed between number of ESE and strength or nature (constitutive vs. alternative) of splice sites. Furthermore, deletion of ESEs has been shown to have no impact on certain constitutive splicing events under optimal conditions (46). We did not observe any ESE-like sequence within the exonic unpaired elements described here. Further research is required to understand possible functional interdependencies between EUE and ESE.

In brief, our work led to the discovery of EUE, which mediates functional interactions of SR proteins to the pre-mRNA, adding to the context of splice sites. It is highly conceivable that in addition to EUE, there are other features of spliceosomal components that provide context to splice sites, towards fine-tuning appropriate temporal and spatial recognition of splice sites.

AUTHOR CONTRIBUTION

KS designed and performed majority of the experiments, analyzed the data, and wrote the paper. TB designed some experiments and wrote the paper. MF performed protein purification, cloning, and site-

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

directed mutagenesis experiments, and critiqued the paper. RS and WE performed and analyzed the icSHAPE data and wrote the paper. GG designed some experiments, wrote the paper, and supervised the project.

ACKNOWLEDGEMENT

The authors acknowledge Stephan Lueng and Suhyung Cho for some early work, Dr. Xiang-Dong Fu for mammalian expression plasmids, Dr Gene Yeo and Dr Stefan Aigner for critical reading of an earlier version of the manuscript, and Dr Matt Daugherty and Dr Shankar Subramaniam for critical discussion on bioinformatic analyses. This work was supported by NIH grant GM 085490 to GG.

REFERENCES

1. Fredericks AM, Cygan KJ, Brown BA, & Fairbrother WG (2015) RNA-Binding Proteins: Splicing Factors and Disease. *Biomolecules* 5:893-909.
2. Li YI*, et al.* (2016) RNA splicing is a primary link between genetic variation and disease. *Science* 352:600-604.
3. Diederichs S*, et al.* (2016) The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. *EMBO Mol Med* 8:442-457.
4. Will CL & Luhrmann R (2011) Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* 3:pii: a003707.
5. Bentley DL (2014) Coupling mRNA processing with transcription in time and space. *Nat Rev Genet* 15:163-175.
6. Voong LN*, et al.* (2016) Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell* 167:1555–1570.
7. Fong N*, et al.* (2014) Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev* 28:2663-2676.
8. Wong JJ*, et al.* (2017) Intron retention is regulated by altered MeCP2-mediated splicing factor recruitment. *Nat Commun* 8:15134.
9. Pan T & Sosnick T (2006) RNA folding during transcription. *Annu Rev Biophys Biomol Struct* 35:161-175.
10. Jimeno-González S*, et al.* (2015) Defective histone supply causes changes in RNA polymerase II elongation rate and cotranscriptional pre-mRNA splicing. *Proc Natl Acad Sci U S A* 112:14840-14845.
11. Lin CL, Taggart AJ, & Fairbrother WG (2016) RNA structure in splicing: An evolutionary perspective. *RNA Biol* 13:766-771.
12. Shen M*, et al.* (2013) Pyrvinium pamoate changes alternative splicing of the serotonin receptor 2C by influencing its RNA structure. *Nucleic Acids Res* 41:3819-3832.
13. Hiller M, Zhang Z, Backofen R, & Stamm S (2007) Pre-mRNA secondary structures influence exon recognition. *PLoS Genet* 3:e204.

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

14. Buratti E & Baralle FE (2004) Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol Cell Biol* 24:10505-10514.
15. Jin Y, Yang Y, & Zhang P (2011) New insights into RNA secondary structure in the alternative splicing of pre-mRNAs. *RNA Biology* 8:450-457.
16. Lu Z*, et al.* (2016) RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* 165:1267–1279.
17. Ding Y*, et al.* (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* 505:696-700.
18. Spitale RC*, et al.* (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* 519:486-490.
19. Corley M*, et al.* (2017) An RNA structure-mediated, posttranscriptional model of human α-1-antitrypsin expression. *Proc Natl Acad Sci USA* 114:E10244–E10253.
20. Mahen EM, Watson PY, Cottrell JW, & Fedor MJ (2010) mRNA secondary structures fold sequentially but exchange rapidly *in vivo*. *PLoS Biol* 8:e1000307.
21. Taliaferro JM*, et al.* (2016) RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. *Mol Cell* 64:294-306.
22. Lambert N*, et al.* (2014) RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell* 54:887-900.
23. Zhou Z & Fu X-D (2013) Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* 122:191-207.
24. Sapra AK*, et al.* (2009) SR protein family members display diverse activities in the formation of nascent and mature mRNPs *in vivo*. *Mol Cell* 34:179-190.
25. Krainer AR, Maniatis T, Ruskin B, & Green MR (1984) Normal and mutant human β-globin pre-mRNAs are faithfully and efficiently spliced *in vitro*. *Cell* 36:993-1005.
26. Bennett M, Michaud S, Kingston J, & Reed R (1992) Protein components specifically associated with prespliceosome and spliceosome complexes. *Genes Dev* 6:1986-2000.
27. Low JT & Weeks KM (2010) SHAPE-directed RNA secondary structure prediction. *Methods* 52:150-158.
28. Darzacq X*, et al.* (2007) *In vivo* dynamics of RNA polymerase II transcription. *Nat Struct Mol Biol* 14:796-806.
29. Schwanhäusser B*, et al.* (2011) Global quantification of mammalian gene expression control. *Nature* 473:337-342.
30. Mathews DH (2014) RNA secondary structure analysis using RNAstructure. *Curr. Protoc. Bioinformatics* 46:12.16.11-12.16.25.
31. Fu X-D (1993) Specific commitment of different pre-mRNAs to splicing by single SR proteins. *Nature* 365:82-85.
32. Sanford JR*, et al.* (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res* 19:381-394.
33. Pandit S*, et al.* (2013) Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol. Cell* 50:223-235.
34. Van Nostrand EL*, et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* 13:508-514.
35. Zhu J & Krainer AR (2000) Pre-mRNA splicing in the absence of an SR protein RS domain. *Genes Dev* 14:3166–3178.
36. Hertel KJ & Maniatis T (1999) Serine–arginine (SR)-rich splicing factors have an exon-independent function in pre-mRNA splicing. *Proc. Natl. Acad. Sci. USA* 96:2651–2655.
37. Das R & Reed R (1999) Resolution of the mammalian E complex and the ATP-dependent spliceosomal complexes on native agarose mini-gels. *RNA* 5:1504-1508.

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. Staknis D & Reed R (1994) SR proteins promote the first specific recognition of pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol. Cell Biol.* 14:7670-7682.
39. Bratt E & Ohman M (2003) Coordination of editing and splicing of glutamate receptor pre-mRNA. *RNA* 9:309-318.
40. Kawaguchi R & Kiryu H (2016) Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome. *BMC Bioinformatics* 17:203.
41. Nelson KK & Green MR (1988) Splice site selection and ribonucleoprotein complex assembly during in vitro pre-mRNA splicing. *Genes Dev* 2:319-329.
42. Screaton GR*, et al.* (1995) Identification and characterization of three members of the human SR family of pre-mRNA splicing factors. *EMBO J* 14:4336-4349.
43. Wang J, Smith PJ, Krainer AR, & Zhang MQ (2005) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res* 33:5053-5062.
44. Zhang XH-F & Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes & Dev.* 18:1241-1250.
45. Fairbrother WG, Yeh R-F, Sharp PA, & Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* 297:1007-1013.
46. Schaal TD & Maniatis T (1999) Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol. Cell Biol.* 19:261-273.
47. Cho S*, et al.* (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.
48. Tacke R, Chen Y, & Manley JL (1997) Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer. *Proc Natl Acad Sci USA* 94:1148-1153.
49. Chandler SD, Mayeda A, Yeakley JM, Krainer AR, & Fu X-D (1997) RNA splicing specificity determined by the coordinated action of RNA recognition motifs in SR proteins. *Proc. Natl. Acad. Sci. USA* 94:3596-3601.
50. Schindelin J*, et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nature Methods* 9:676-682.
51. Darty K, Denise A, & Ponty Y (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25:1974-1975.
52. Harmanci A, Sharma G, & Mathews DH (2007) Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *BMC Bioinformatics* 8:130.
53. Mayeda A & Krainer AR (1999) Preparation of Hela cell nuclear and cytosolic S100 extracts for *in vitro* splicing. in *Methods in Molecular Biology (RNA-protein interaction protocols)*, ed Haynes SR (Humana Press (Springer)), pp 309-314.
54. Zubradt M*, et al.* (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing *in vivo*. *Nat Methods* 14:75-82.

Methods

Cloning and protein expression

cDNAs for different SR proteins were cloned from purchased from Addgene or Open Biosystems. The cDNAs were cloned in T7 promoter-based *E. coli* expression vectors and were expressed as hexa-

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

histidine fusion proteins. Fusion proteins contain a *TEV* protease cleavage site for removal of the tag. Tag-removed proteins were used in all experiments. Proteins were expressed in *E. coli* BL21 (DE3) cells with isopropyl β-D-1-thiogalactopyranoside induction, and purified by Ni-NTA (Ni^{2+}-nitrilotriacetate) affinity chromatography. The His_6-tags were removed by treatment with His_6-TEV protease overnight at room temperature and then uncleaved proteins were separated by passing through respective resins. Untagged proteins were further purified by either size-exclusion (Superdex 75; GE Healthcare Lifesciences) or cation exchange chromatography (SP sepharose or Mono S; GE Healthcare Lifesciences). Three functional but modified variants of SR proteins were used: RNA binding domain of SRSF1 (1-203 a.a.)(47), RNA binding domain of SRSF5 (1-184 a.a.)(48), and RNA binding domain of SRSF2 (1-127 a.a.) in chimera with fully phosphomimetic serine-arginine domain SRSF1 (197-246 a.a.) with all serines replaced with glutamate (49). All purified proteins were confirmed to be RNase-free by incubating a small aliquot of the purified protein with a long RNA overnight at room temperature and analyzing the RNA quality by urea PAGE after phenol extraction.

Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE)

25 nM denatured and renatured *β-globin* pre-mRNA was incubated with 250 nM SRSF1-RBD or equal volume SRSF1 protein storage buffer under EMSA conditions for 15 min; similarly, denatured and renatured 25 nM *AdML* pre-mRNA was incubated with 250 nM SRSF1. Freshly prepared 2 mM (final concentration) N-Methylisatoic anhydride (NMIA, Aldrich) in dimethyl sulfoxide (DMSO, Sigma) was added to the test reaction, equal volume of DMSO was added to the control reaction and the reaction was allowed to proceed for 15 min at 30 ^{o}C and then 15 min at room temperature. Next, the reaction was diluted in proteinase K buffer and protein was digested with proteinase K (New England Biolabs) before the RNA was purified. Reverse transcription reaction was carried out using MuLV Super-RT reverse transcriptase (Biobharati Life Science, India) using manufacturer's protocol at 50°C and the products were run in 12 % and 8 % 7 M urea sequencing gel of 0.35 mm thickness. Each RT product was resolved up to two different lengths on 8 % gel. By this method, up to 125 nucleotides could be

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

resolved under each primer. The individual bands were identified and quantified using Fiji (50). The values were processed as described before (27). Briefly, all numbers in a series under each primer was divided by the average of the top 10% values excluding the outliers. Values 1.5 times of the median value of each series (capped at top 5% as the number of bands under each primer was small, about 125-150) were considered outliers. Using the processed values as pseudo-free energy, the RNA was folded using 'RNAstructure' (30) at 298.15 K with a slope and intercept of 2.6 and -0.8, respectively. All negative values were considered to be zero. Maximum values were capped at 3.0. The RNA secondary structure models were drawn using VARNA (51). The following primers were used for reverse transcription of *β-globin* – 5′ACGTGCAGCTTGTCACAGTG (*βgRT1*), 5′TTTCTTGCCATGAGCCTTC (*βgRT2*), 5′AGTGGACAGATCCCCAAAG (*βgRT3*), 5′GGAAAATAGACCAATAGGC (*βgRT4*), 5′TTCTCTGTCTCCACATGCC (*βgRT5*), 5′AACTTCATCCACGTTCACC (*βgRT6*). The final primer was replaced with the following for reverse transcription of EH3+4 mutant of *β-globin* 5′CCAACTTCATGCTGGTGACG (*EH-RT6*). Reverse transcription of *AdML* was carried out with the following primers – 5′AAGAGTACTGGAAAGACCGC (*AdRT1*), 5′GGACAGGGTCAGCACGAAAG (*AdRT2*), 5′TCGAGGGCAGAAGTCATGCC (*AdRT3*), 5′CAGCGATGCGGAAGAGAGTG (*AdRT4*).

Pre-mRNA constructs

We carried out all our experiments with pre-mRNAs containing constitutively spliced exons and used the following pre-mRNA constructs (exon-intron-exon lengths are given in parenthesis): *β-globin* IVS1 (109+130+204), *AdML* (65+123+50), *SRSF1* IVS1 (403+416+185), *SNRP70* IVS1 (498+768+157), *ACTA2* IVS5 (85+394+162), *HBA1* IVS1 (161+117+205), and SRSF7 IVS2 (181+308+177). The last five constructs were amplified from genomic DNA of HEK293T cells and were cloned into a mammalian transfection vector under CMV promoter.

Predictive structural modeling of the intron-proximal stem-loop of pre-mRNAs using Dynalign:

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Based on the SHAPE-derived secondary structure model of *β-globin* and *AdML*, we anticipated that the intron-proximal stem-loop can be modeled *in silico* using sequence spanning ~ 70-nt upstream and 10-nt downstream of 5′SS as the input sequence for RNA-structure prediction algorithm. We obtained these models of the RNAs using Dynalign (52), a pairwise structure prediction program. Pairwise structure prediction is known to improve accuracy of the predicted structure and hence, we predicted the secondary structures of the segments in comparison with the segment spanning -70-nt through +10-nt of *β-globin* and -65-nt through +10-nt of *AdML* separately. Only those models of the pre-mRNAs in which the intron-proximal stem-loop contained at least -50-nt through +10-nt around the 5′SS and had the lowest energy of folding were selected.

In vivo splicing assay

HeLa cells grown in six well plates (9 cm^2 surface area) were transfected with 1 µg of purified plasmid containing WT or mutant pre-mRNA sequence. RNA was harvested at 36 h after transfection. cDNA was synthesized in 20 µl volume with 100 ng RNA using MuLV Super RT reverse transcriptase (Biobharati Life Science, India) following manufacturer's instructions at 50 ^oC. 1 µl cDNA was used in a 25 µl PCR mix. Sub-saturating PCR products were analyzed on 2 % agarose gels.

Splicing assay and spliceosome assembly assay

Nuclear extract was prepared from *HeLa* cells as published (53). E-complex assembly assay was carried out as described before (37). *In vitro* splicing assay of WT and mutant *β-globin* was carried out as described before (49).

RNA-seq data analysis

IcSHAPE data was analyzed from published available data (18). All exon-exon and exon-intron boundaries were identified using annotations from the GENCODE release M12. Only transcripts annotated to contain retained introns were searched for exon-retained intron boundaries. To identify the position of the retained intron, exon start and stop positions from retained intron transcripts were

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compared to those of coding transcripts originating from the same gene. Exons from retained intron transcripts which spanned two sequential exons in a related coding transcript were considered to contain introns, and the exon-intron boundary was defined as the 3′ end of the upstream exon in the coding sequence. A 70 bp range upstream of this boundary was extracted for analysis. For comparison, 70 bp of sequence upstream of 10,000 randomly selected annotated exon-exon junctions in coding transcripts were analyzed. For both sets, transcripts with another exon-exon or exon-intron boundary within the 70 bp were discarded.

In vivo DMS-seq

HeLa cells transfected with the plasmids were treated with dimethyl sulfate as described before (54). Nuclei were isolated by the following method. Transfected and DMS-treated cells were washed with chilled PBS. Scraped cells were centrifuged for 30 s at 1000 g, supernatant was removed, cells were resuspended in 10 mM HEPES 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 0.05% Nonidet P-40, and kept on ice for 10 min, with occasional flicking. Cells were centrifuged for 10 min at 730 g at 4 °C. The pellet was resuspended in Trizol (Thermoscientific) and proceeded with RNA isolation with manufacturer's protocol. Reverse transcription of *β-globin* constructs was carried out with *βgRT5* and that of *SRSF1* IVS1 was carried out with 5'TCTAAGAGCCCCCGCTTCCC3', a primer nested in the intron, 82-nt downstream of the 5'ss. Reverse transcription and densitometric analyses were carried out as in SHAPE.

FIGURE LEGENDS

Fig. 1: Secondary structure model of *β-globin* and its bearing on splicing. (A) SHAPE-derived secondary-structure model of *β-globin*; nucleotides are color-coded according to their SHAPE-reactivity (see associated legend); sites of exonic hybridization (EH), exonic de-hybridization (ED), exonic deletion (ΔL), and intronic hybridization (IH) are indicated on the map; original and mutated nucleotide

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequences are shown at the bottom. (B) Secondary structure models of *β-globin* 5' exon in EH, ΔL, and ED mutants, predicted based on SHAPE-derived secondary structure model of the WT *β-globin* 5' exon as shown in (A). (C) Splicing products of WT, EH, EH+ED, and ED mutants of *β-globin* in transfection-based splicing assay; quantification of properly spliced mRNAs as a percentage of total RNA generated from the transfected plasmids are shown as %mRNA using the formula: (mRNA/total RNA)/(WT mRNA/total WT RNA)*100. (D) Splicing products obtained by *in vitro* splicing assay with radiolabeled WT *β-globin*, and its EH3+4 mutant; asterisks indicate aberrantly spliced products.

Fig. 2: The exonic unpaired elements do not show conservation or requirement of specific nucleotide sequence: (A) Secondary structure model of protein-free *β-globin* EH3+4 mutant obtained by *in vitro* SHAPE. (B) Putative secondary structure models of the intron-proximal stem-loop in 5′ exon of *β-globin* mutants EH1+2+3+4, EH1+2+3+4+TL, and EH1+2+3+4+5L[2]; the original and mutated nucleotide sequences are shown at the bottom; the cryptic spliced product obtained from EH1+2+3+4 as indicated with an * in lane 2 of (C) is indicated by two red borders. (C) Transfection-based splicing assay of EH1+2+3+4, EH1+2+3+4+5L2, and EH1+2+3+4+TL mutants of *β-globin* (lanes 1-4); transfection-based splicing assay of substitution mutants of *β-globin* as shown in (D) (lanes 5-7). (D) Sites and nucleotide sequences of substitution mutations of *β-globin*. (E) Sequence logo obtained from the alignment of 102 exon-intron junctions (from human chromosome 1) around 5′SS between -70 and +10-nt showing a dearth of nucleotide conservation in this region.

Fig. 3: *In vivo* structural modulation of transfected WT and mutant *β-globin* analyzed by DMS-seq: (A) Scheme of *in vivo* DMS-seq analysis. (B) Quantified reverse transcription stops (DMS-reactivity) of individual nucleotides (A's & C's only) of WT *β-globin* (o) and its EH3+4 mutant (Δ) are plotted against nucleotide positions; dot-bracket notations, which is aligned with the nucleotide positions in the plot, of the secondary structure model of protein-free WT *β-globin* and its EH3+4 mutant are

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

shown at the bottom. (C) DMS-reactivity of individual nucleotides (A's & C's only) of endogenous *SRSF1* IVS1 is plotted against nucleotide positions; dot-bracket notations of both predicted structural models of *SRSF1 IVS1* as shown in Fig. S2 are given at the bottom.

Fig. 4. Exonic unpaired elements induce SR protein-mediated pre-mRNA structural modulation: (A) Genome browser views of 5′ exon of *SRSF1* IVS1 with mapped SR protein eCLIP reads (for SRSF1, SRSF7, and SRSF9) showing that the exonic region immediately upstream of 5′SS overlaps with high confidence SR protein binding sites; mapped SR protein read densities from two eCLIP biological replicate experiments performed in HepG2 cells and one size-matched input are plotted at the same y-axis scale; scale bar, and the directions of transcription by an arrow are denoted (www.encodeproject.org). (B) (Top) SHAPE reactivity differential (ΔSHAPE reactivity) of free RNA and SRSF1-bound RNA are shown for WT *β-globin* (blue line) and its EH3+4 mutant (red line). (Bottom) SHAPE reactivity differential of free RNA and SRSF2-bound RNA are shown for WT *β-globin* (blue line) and its EH3+4 mutant (red line); dot-bracket notations of WT and mutant *β-globin* are shown. (C) E-complex assembly assay with WT *β-globin* and its EH3+4 mutant with ATP-depleted nuclear extract.

Fig. 5. Analysis of icSHAPE data from mouse embryonic stem cells: (A & B) Scheme of *in vivo* and *in vitro* icSHAPE analysis. (C) *In vitro* icSHAPE data for 70-nt region upstream of 5′SS in 10000 exon-exon junctions (circle) and in 234 exon-intron junctions from retained introns (square) in polyadenylated mRNAs. (D) *In vivo* icSHAPE data for 70-nt region upstream of 5′SS in 10000 exon-exon junction (circle) and in 291 exon-intron junctions from retained introns (square) in polyadenylated mRNAs; of the exon-intron junctions, 219 are shared by both datasets. (E) Box-plot analyses showing distribution of free energy of folding; blue plot indicates free energy of folding of the target regions (+10 through -70 around 5′SS) of 1000 5' UTR exons obtained from 10 human chromosomes; orange plot indicates the same of the non-target regions (-71 through -150) of the same set of exons; grey plot indicates the same of the

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

target regions (+10 through -70 around 5′SS) of the 291 exons followed by retained introns (RI) as shown in Fig. 5D; 'x' indicates the mean value, horizontal bar inside the box the median value, bottom and top sides of the boxes the first and third quartiles, respectively, whiskers variability outside first and third quartiles, and individual dots outliers.

SUPPLEMENTARY FIGURE LEGENDS

Fig. S1: SHAPE-derived secondary structure models of pre-mRNAs and their bearings on splicing: (A) SHAPE-derived secondary-structure model of *AdML*; nucleotides are color-coded according to their SHAPE-reactivity (see associated legend); nucleotides without available SHAPE data are colored black; sites of exonic hybridization (EH), exonic de-hybridization (ED), and intronic hybridization (IH) are indicated on the map; original and mutated nucleotide sequences are shown at the bottom. (B) Schematic showing *AdML* exonic stem-loop hybridization and de-hybridization. (C) Splicing products obtained by transfection-based splicing assay of WT, EH, EH+ED, and ED mutants of *AdML*; positions of pre-mRNA and properly spliced mRNA on agarose gel are indicated in red; quantification of properly spliced mRNAs as a percentage of total RNA generated from the transfected plasmids are shown as %mRNA using the formula: (mRNA/total RNA)/(WT mRNA/total WT RNA)*100; an asterisk indicates aberrantly spliced product. (D) Splicing products obtained by transfection-based splicing assay of WT and IH mutants of *β-globin*.

Fig. S2: Regulation of splicing by exonic unpaired elements in multiple constitutive exons: (top) Two alternative predicted secondary structure models of *SRSF1* IVS1 (403+416+185) and one of each of *SNRP70* IVS1 (498+768+157), *ACTA2* IVS5 (85+394+162), *HBA1* IVS1 (161+117+205), and SRSF7 IVS2 (181+308+177) encompassing -70-nt through +10-nt around 5′SS with the area of hybridization-mutation of exonic loops mapped onto it; the original and mutated sequences are shown. (Bottom) splicing products of WT substrates and their hybridization mutants in transfection-based splicing assay;

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

E=empty vector; an asterisk indicates aberrantly spliced product; +1 indicates the first nucleotide of the intron.

Fig. S3. Bio-informatic analyses of *β-globin* substitution mutants: (A) Secondary structure models of the intron-proximal stem-loops of 5' exon of WT *β-globin*, and its M1+2+3+4+6+7 and M1+2+3+4+6+7+8 mutants as predicted by 'RNAstructure'. (B) Alignment of *β-globin* WT and its M1+2+3+4+6+7 substitution mutant to show extensive substitution in the mutant pre-mRNA.

Fig. S4. SRSF1-mediated structural modulation of *AdML*: SHAPE reactivity differential (ΔReactivity) of protein-free RNA and SRSF1-bound RNA *in vitro* are shown for WT *AdML*; dot-bracket notation of secondary structure model of protein-free *AdML in vitro* (nucleotide numbers are aligned with those in the plot) is shown.

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

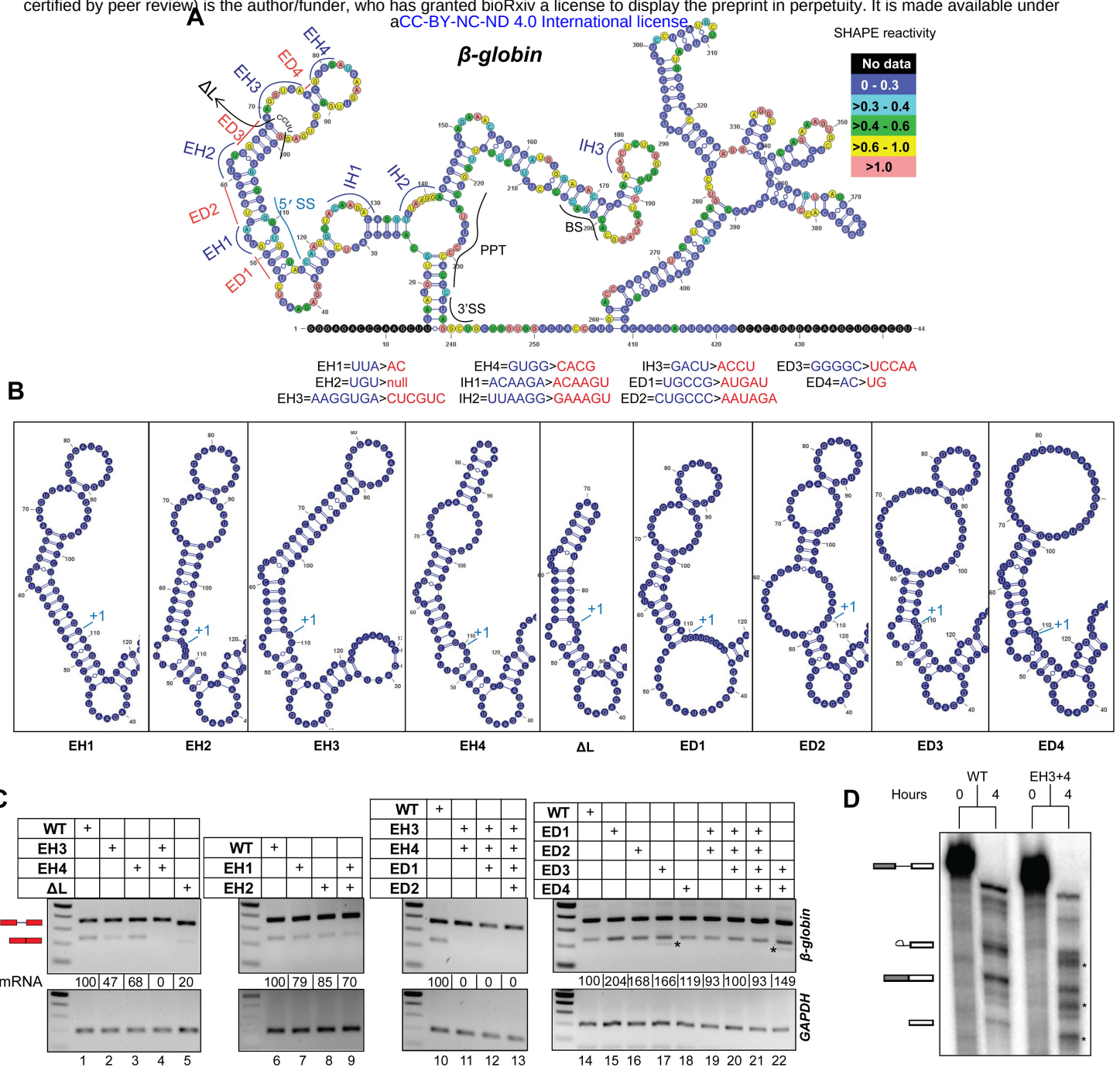


Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

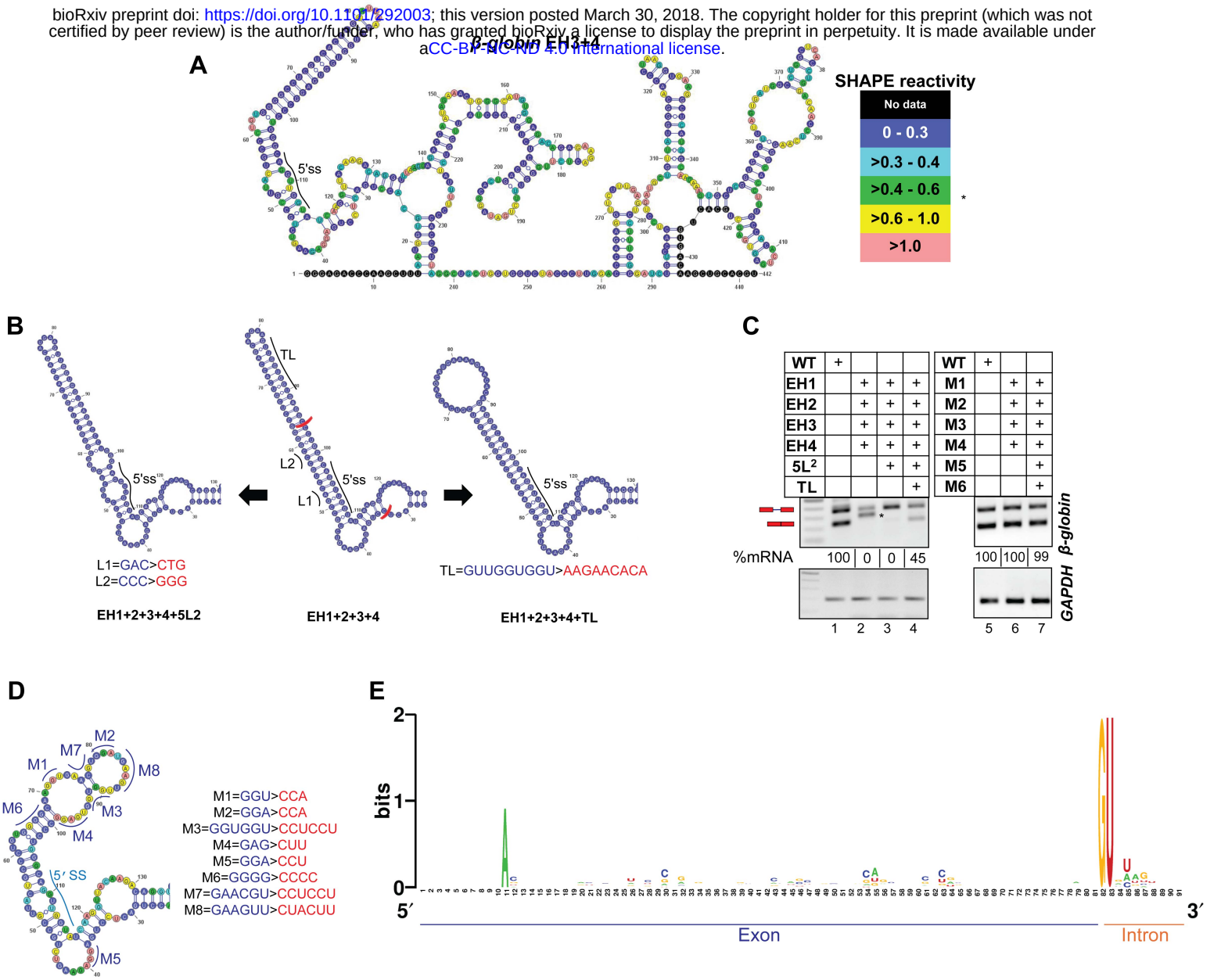


Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

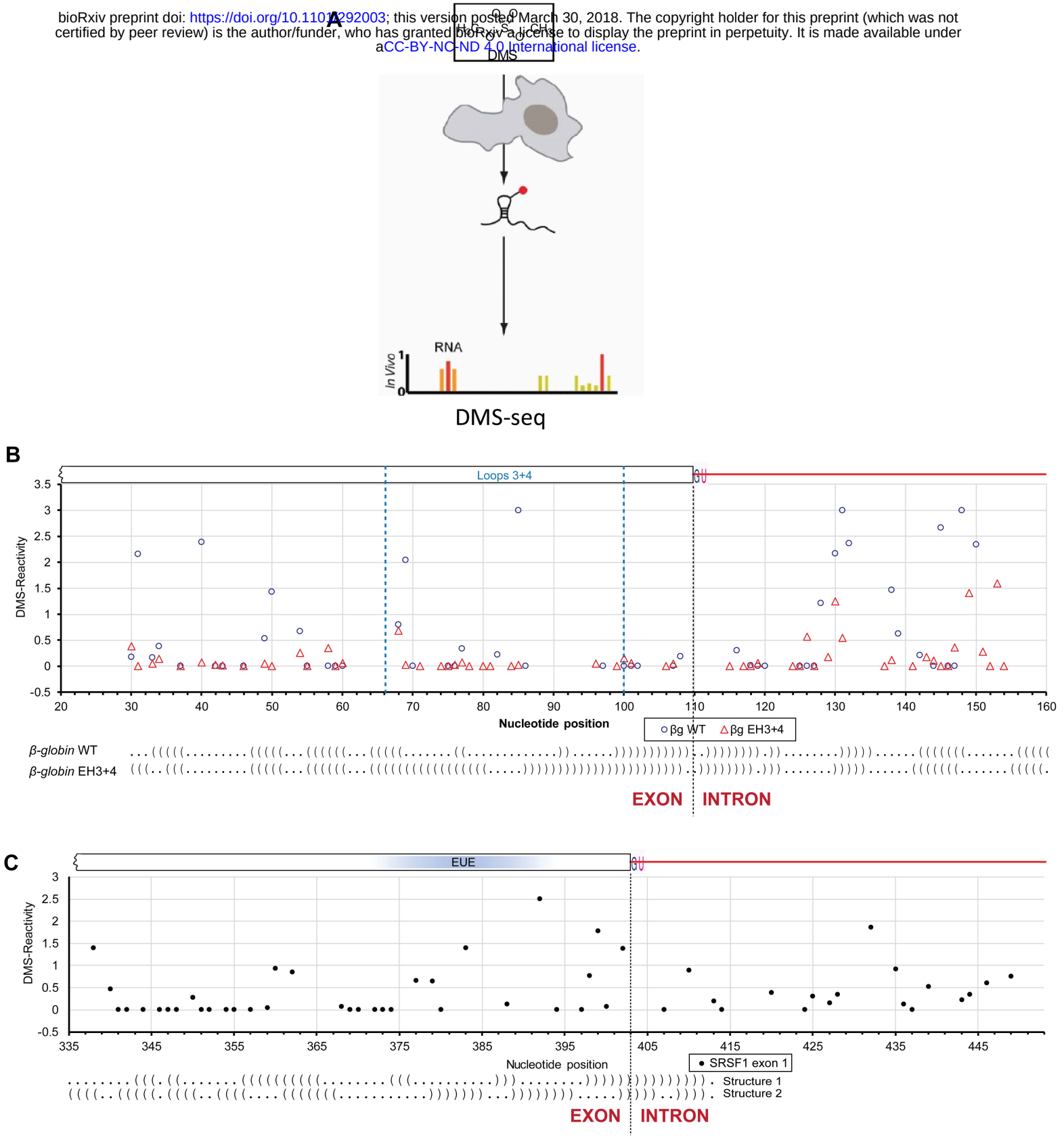
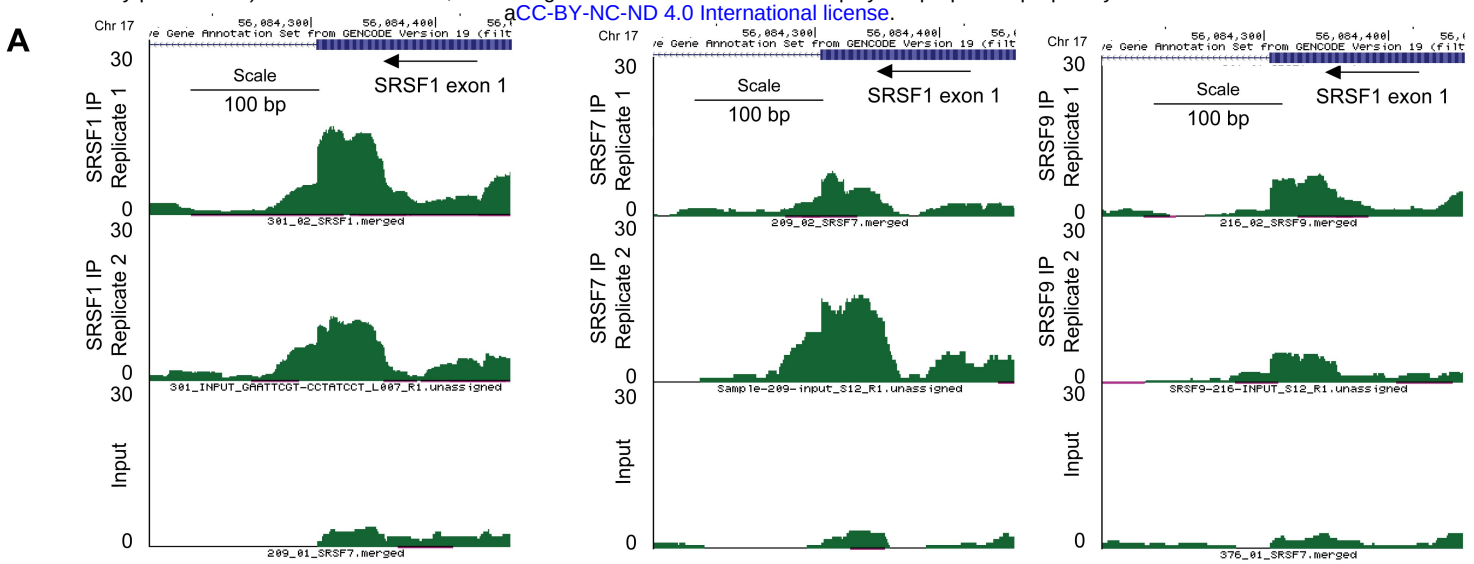
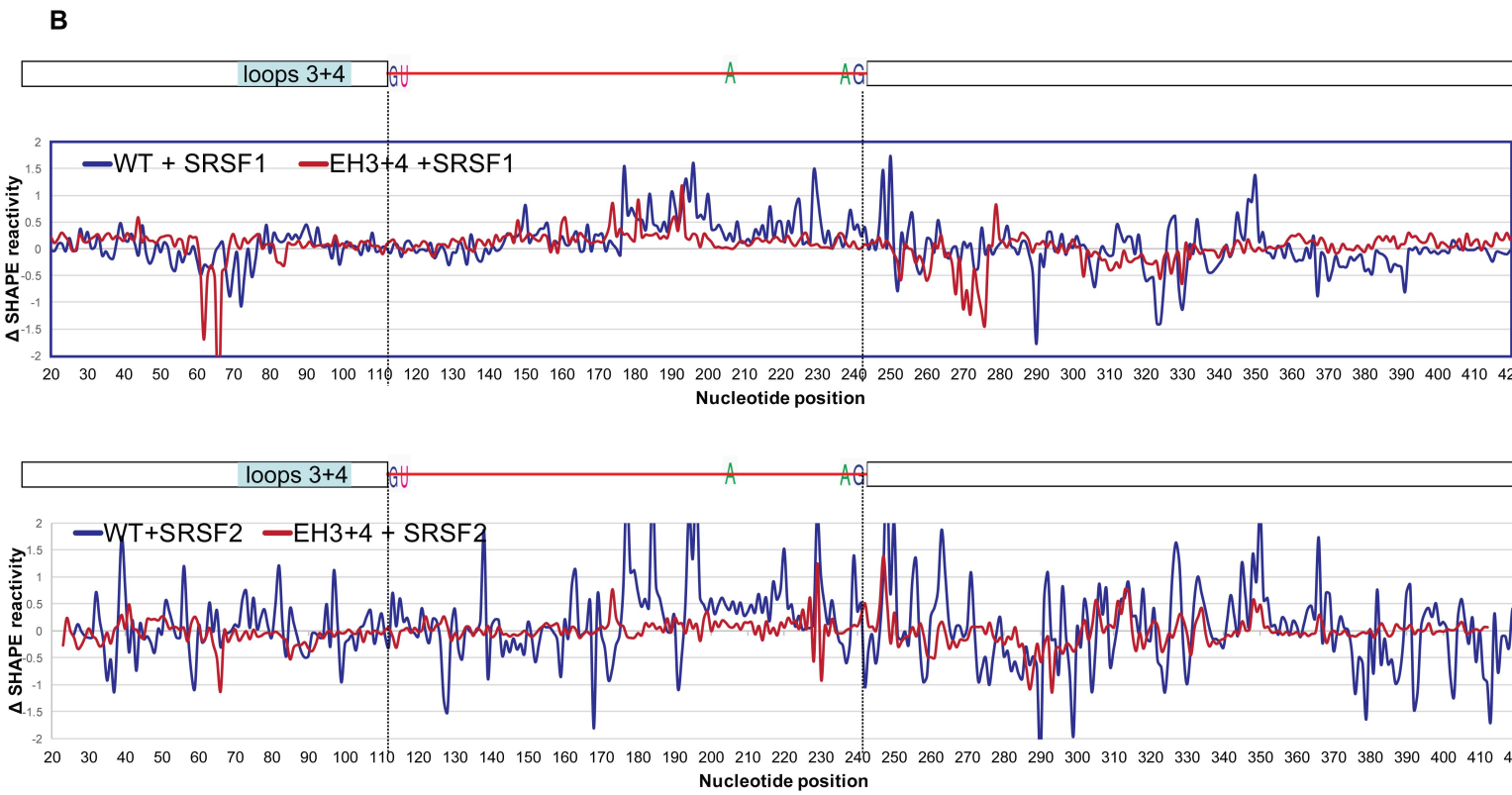


Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.





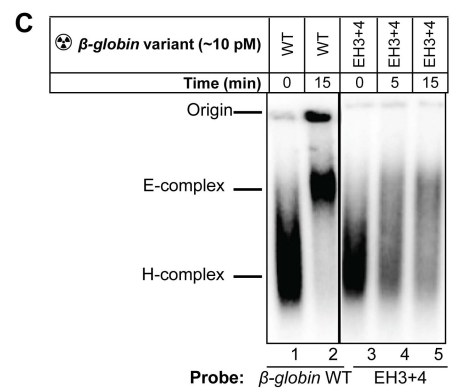


Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

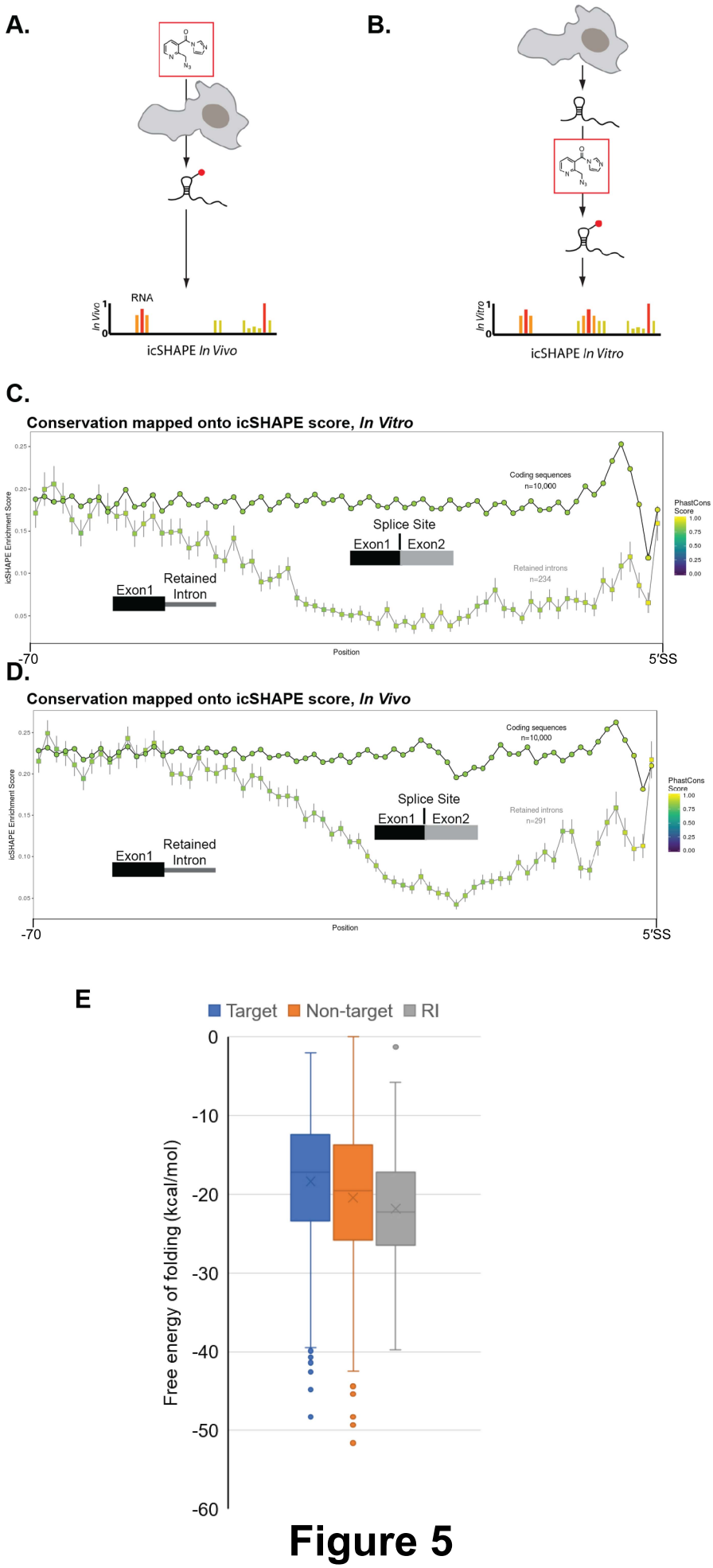


Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

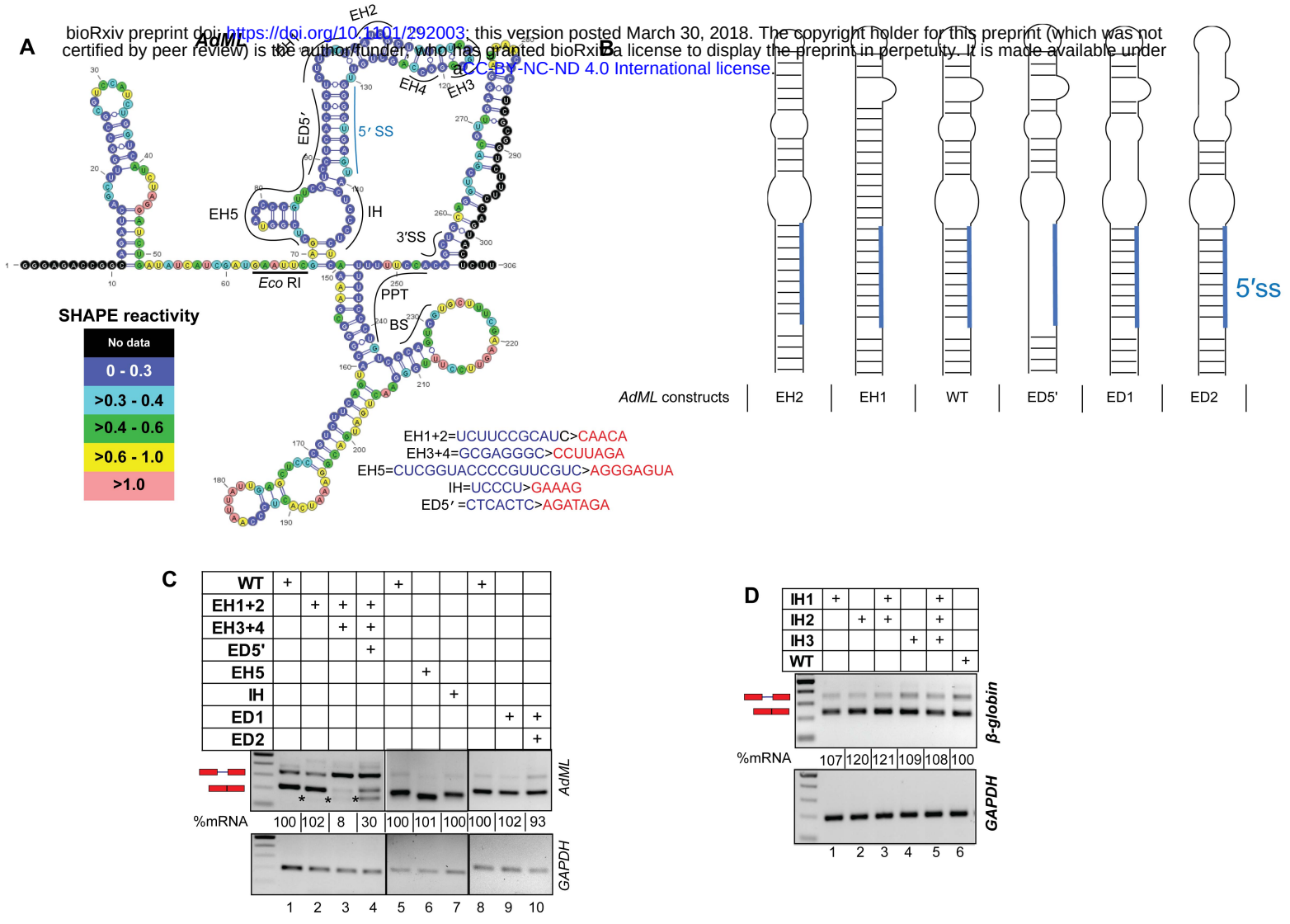


Figure S1

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

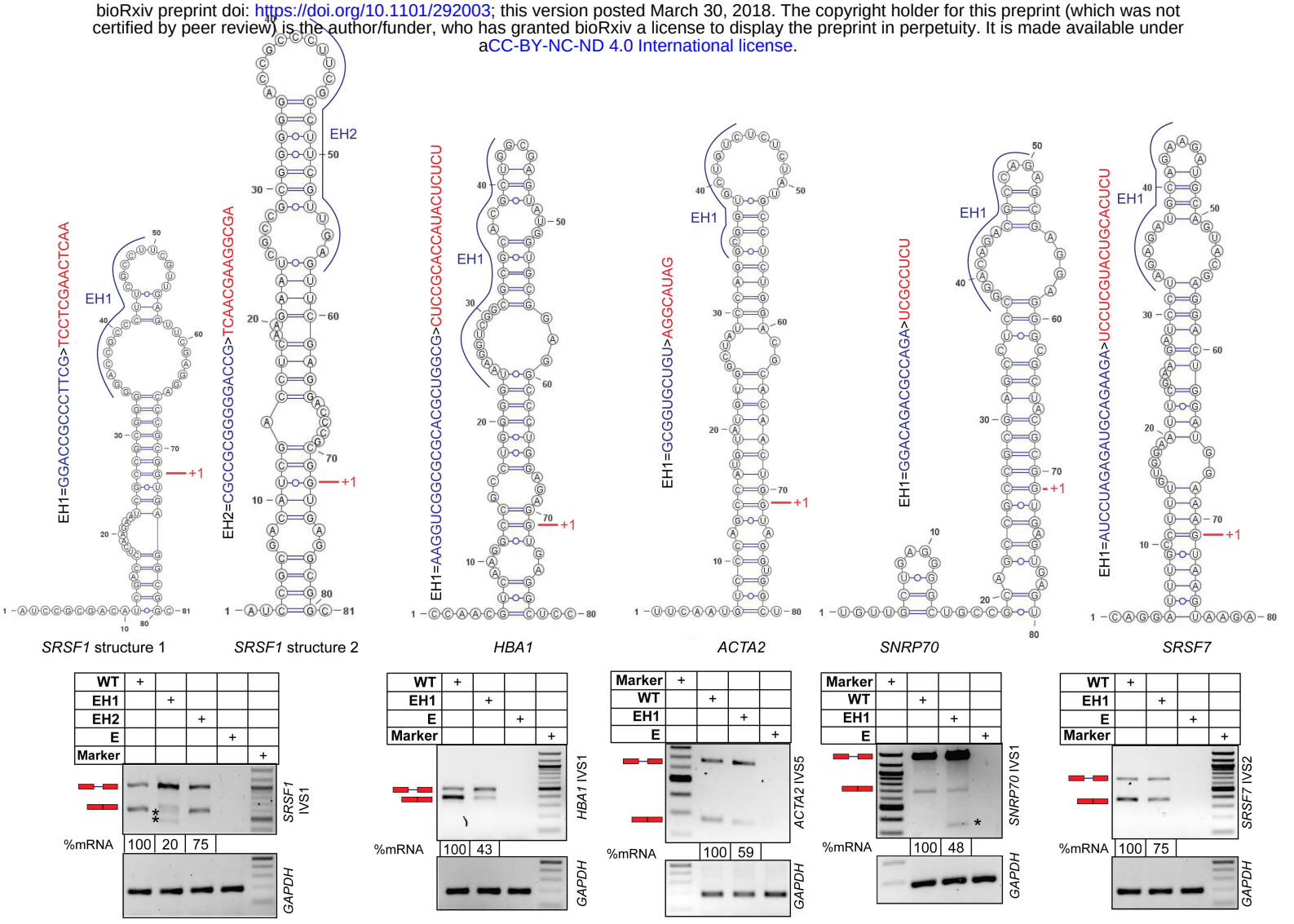
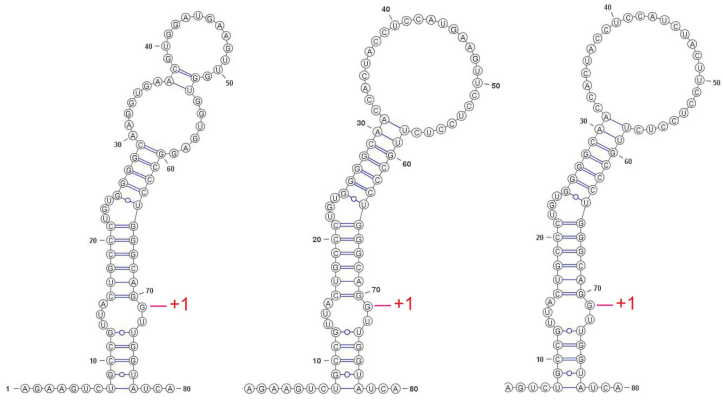


Figure S2

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A



β-globin *β-globin* M1+2+3+4+5+6 *β-globin* M1+2+3+4+5+6+7

B

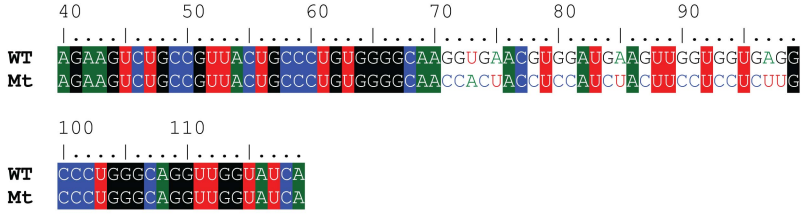


Figure S3

bioRxiv preprint doi: https://doi.org/10.1101/292003; this version posted March 30, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

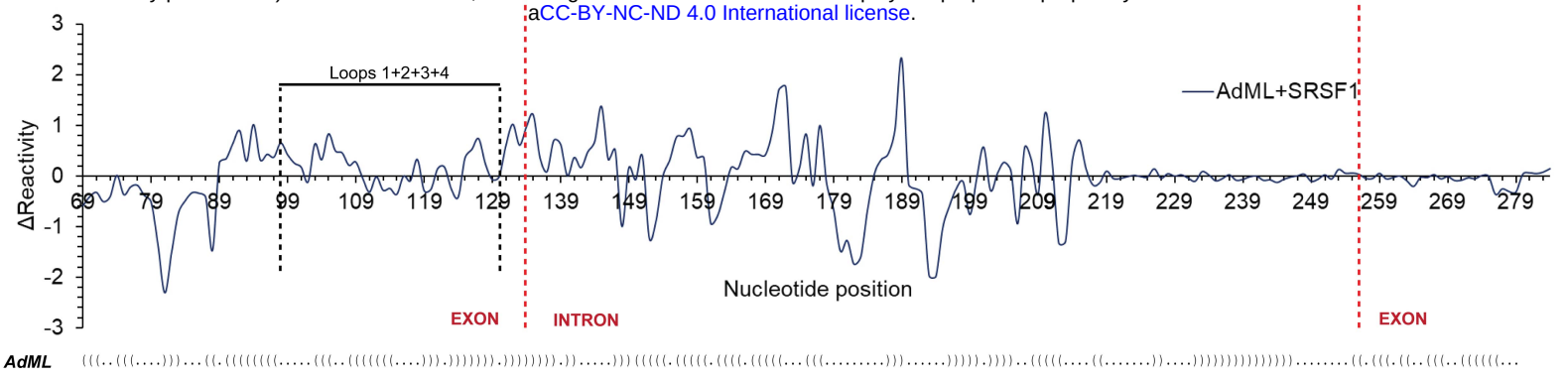


Figure S4