In vitro assembly of an early spliceosome defining both splice sites

Kaushik Saha, Mike Minh Fernandez, Tapan Biswas, Charles Leonard Mallari Lumba, Gourisankar Ghosh*

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

*Correspondence: gghosh@ucsd.edu

ABSTRACT

For splicing of a metazoan pre-mRNA, the four major splice signals – 5′ and 3′ splice sites (SS), branch-point site (BS), and a poly-pyrimidine tract (PPT) – are initially bound by splicing factors U1 snRNP, U2AF35, SF1, and U2AF65, respectively, leading up to an early spliceosomal complex, the E-complex. The E-complex consists of additional components and the mechanism of its assembly is unclear. Hence, how splice signals are organized within E-complex defining the exon-intron boundaries remains elusive. Here we present *in vitro* stepwise reconstitution of an early spliceosome, assembled by cooperative actions of U1 snRNP, SRSF1, SF1, U2AF65, U2AF35, and hnRNP A1, termed here the recognition (R) complex, within which both splice sites are recognized. The R-complex assembly indicates that the SRSF1:pre-mRNA complex initially defines a substrate for U1 snRNP, engaging exons at both ends of an intron. Subsequent 5′SS-dependent U1 snRNP binding enables recognition of the remaining splice signals, defining the intron. This R-complex assembly indicates the minimal constituents for intron definition revealing mechanistic principles behind the splice site recognition.

INTRODUCTION

Splicing of metazoan pre-mRNA requires four *cis*-acting signal elements: 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 3′SS (1). U1 snRNP, SF1, U2AF65, and U2AF35 recognize 5′SS, BS, PPT, and 3′SS, respectively, within early spliceosomal complexes, such as the E-complex. Current knowledge of splice site recognition and E-complex assembly is summarized in Fig. S1.

In E-complex, the 5′SS base pairs with an available nine nucleotide stretch (9-nt; 5′AUACΨΨACC…3′) at the 5′ end of U1 snRNA, the RNA component of U1 snRNP. In case of constitutively spliced introns, the extent of complementarity between the 5′ end of U1 snRNA and 5′SS is roughly 7-nt; however, additional contacts have been proposed to be necessary for a stable interaction of U1 snRNP with the spliceosomal complex (2, 3). The complementarity between 5′SS and U1 snRNA appears to be reduced to only two nucleotides in many cases. Despite structural and biochemical information on the interaction between purified U1 snRNP and a short 5′SS-like RNA fragment (4), the recruitment of U1 snRNP to 5′SS of varied sequences (especially where the base pair complementarity is marginal) within full-length pre-mRNAs is unclear (5). It is strongly likely that the ‘context’ of 5’SS plays a critical role in this recognition, but this context-dependence of 5′SS has not been well established.

The 3′SS is recognized by an intricate network of multiple factors (SF1, U2AF65, and U2AF35) interacting with the 3′ end of the introns. U2AF65 interacts with the PPT, and contacts the BS through its N-terminal RS domain; moreover, its C-terminal UHM domain interacts with the N-terminal ULM domain of SF1 that is bound to the BS. U2AF35 contacts the AG dinucleotide of an authentic 3′SS, and its interaction with the ULM domain of U2AF65 guides U2AF65 to the authentic PPT (Fig. S1). Recent investigations indicate that hnRNP A1 interacts with the U2AF65/U2AF35 heterodimer at the 3′SS, and this ensures appropriate engagement of U2AF65 to PPT followed by an authentic 3′SS (i.e. AG dinucleotide) (6, 7). The spatial context of 3′SS that is recognized within the early spliceosomal complexes is not clear.

Earlier studies indicate that 5′SS and 3′SS need to be in close proximity for efficient assembly of the E-complex (8-10). The major events that bring the two splice sites close to each other initially remain to be identified. Since E-complex consists of a large number of protein components (11), attempts to identify the essential/minimal components that could suffice for an efficient recognition of both splice sites proved to be difficult so far.

Splice site recognition has been shown to be dependent on several members of the serine-arginine-rich (SR) family proteins such as SRSF1 and SRSF2 in both constitutively and alternatively spliced pre-mRNAs (12). All SR proteins contain an N-terminal RNA-binding domain with one or two RNA recognition motifs (RRMs) and a C-terminal Arg-Ser-rich (RS) domain of varying lengths. The molecular mechanism of action of SR proteins in activation of splicing is yet to be unraveled.

In the accompanying manuscript, we have shown that exonic unpaired elements (EUE) immediately upstream of 5′SS mediates pre-mRNA structural modulation by SR proteins, a process essential for E-complex assembly. In the current report, we show that the pre-mRNA structural modulation creates a substrate for U1 snRNP. Binding of U1 snRNP in a 5′SS-dependent manner to this pre-mRNA:SR protein complex enables presentation of the BS, PPT, and 3’SS for recognition by SF1, U2AF65, and U2AF35. The assembly of these factors on the pre-mRNA in splice signal dependent manner leads to formation of the recognition (R) complex, analysis of which provides mechanistic insights into the process of splice site recognition.

RESULTS:

The splicing-conducive SRSF1:pre-mRNA complex is a substrate for U1 snRNP binding

We have previously reported that SR protein-mediated pre-mRNA structural modulation effectuates the assembly of the E-complex (the accompanying manuscript). Since the E-complex has always been assembled with the nuclear extract, it is unclear whether there exist other intermediates *en route* to the E-complex formation. To analyze the steps of splice site recognition with purified components *in vitro*, we assembled and purified U1 snRNP, and analyzed its homogeneity and integrity by SDS-PAGE and negatively stained EM imaging (Fig. S2A, S2B). For U1 snRNP reconstitution, a truncated variant of SNRP70 (SNRP70 ΔRS, residues 1-215) was used in all cases. We also prepared another particle where the C-terminus of SNRPB subunit was truncated (SNRPB$_{174}$, residues 1-174) generating U1 snRNP B$_{174}$. These two variants of the particle – U1 snRNP and U1 snRNP B$_{174}$ – were used interchangeably in our experiments. Removal of the RS domain of SNRP70 has been shown not to affect the viability of *Drosophila* (13). Deletion of the C-terminal R-rich domain of SNRPB does not affect the structural integrity of U1 snRNP (14). Therefore, we anticipated the behavior of these two variants of U1 snRNP to be close to that of the WT U1 snRNP. For SRSF1, we used SRSF1-RBD (residues 1-203), which is a functional truncated variant of SRSF1 (15), unless otherwise indicated. U1 snRNP showed no detectable binding to *β-globin* suggesting masked 5′SS (Fig. 1A, lanes 14-16). EMSA shows that U1 snRNP forms complexes with the radiolabeled *β-globin* pre-mRNA only in the presence of SRSF1-RBD, (Fig. 1A, lanes 3-13). The progressively greater mobility of the pre-mRNA-complex with increasing concentration of U1 snRNP indicates competitive displacement of some SRSF1 molecules. U1 snRNP B$_{174}$ also formed complexes with *β-globin* only in presence of SRSF1 (Fig. 1B, compare lanes 4 and 6). Both anti-SRSF1 and anti-SNRPC antibodies super-shifted the U1 snRNP-dependent complexes indicating the presence of both SRSF1 and U1 snRNP/U1 snRNP B$_{174}$ in the U1 snRNP-dependent complexes (Fig. 1C, lanes 9, 18, 20). We also showed that U1 snRNP recruitment does not occur in *AdML* too in the absence of SRSF1 (Fig. S2C). These results indicate that SRSF1-mediated modulation of the pre-mRNA secondary structure, and assembly of the splicing-conducive pre-mRNA complex creates the substrate for U1 snRNP binding. Pre-mRNA complexes formed with full-length SRSF1 with fully phosphorylated-mimetic RS domain with all 18 serine residues replaced with glutamate (SRSF1-RE) (16) showed similar results as SRSF1-RBD (Fig. S2D, compare lanes 3 & 5);

as negative controls, we also used non-functional variants of full-length SRSF1, SRSF1-RERA (hypophosphorylated mimetic full-length SRSF1 with only 12 serine residues of the RS domain replaced with glutamate) (16) and SRSF1-RS (i.e. WT variant), neither of which could recruit U1 snRNP (Fig. S2D, compare lanes 5, 7, & 9). However, because the complexes assembled with SRSF1-RE did not resolve well in the native gel used for EMSA, we used unlabeled full-length *β-globin* with three MS2 binding loops at the 3′ end and assembled the complex with concentrations indicated in Fig. S2D. MBP-tagged MS2 coat protein was used to pull down the RNA after assembly of the complex. Consistent with the EMSA results, strong interaction between the pre-mRNAs and clear U1 snRNP protein bands were observed only in the presence of SRSF1-RE (Fig. 1D).

Additional SR proteins lower the SRSF1 level required for U1 snRNP recruitment

Although results obtained thus far demonstrate that SRSF1-mediated structural modulation of the pre-mRNA is essential for U1 snRNP recruitment, they do not uncouple the role of SRSF1 in stabilizing U1 snRNP on the pre-mRNA (17, 18) from its function of splicing-conducive pre-mRNA-complex assembly. Therefore, we examined if other SR proteins, SRSF2 and SRSF5 recruit U1 snRNP. Both SR proteins promote the splicing of *β-globin* (19). SHAPE experiment with SRSF2 also indicated that this SR protein modulates the secondary structure of the *β-globin* (the accompanying manuscript). In contrast to SRSF1, SRSF2 or SRSF5 failed to recruit U1 snRNP on its own (Fig. 2A, compares lanes 5, 9, 13). This indicates that SRSF1, in addition to assembling a splicing-conducive pre-mRNA complex, also stabilizes U1 snRNP on the pre-mRNA.

Binding of exons by SR proteins could exhibit compensatory nature, where depletion of one SR proteins may be compensated for by another SR protein for exon binding; occasionally, the occurrence of compensatory binding changes the splicing outcome, as does lack of it, as an essential mechanism of splicing regulation (20). Therefore, we examined whether three SR proteins, SRSF1, SRSF2, and SRSF5, all of which are known to promote splicing of *β-globin* (19), exhibit compensatory binding, and whether their compensatory binding recruits U1 snRNP. Fig. 2B shows that while 60 nM, but not 20 nM, SRSF1 alone is capable to form stable *β-globin*-U1 snRNP complex, 10 nM SRSF1 forms stable *β-globin*:U1 snRNP complexes in presence of 6 nM SRSF2 plus 6 nM SRSF5 or 12 nM of either (compare lanes 3 & 7-13). We verified this observation by purifying the *β-globin*:SR:U1 snRNP complex formed with both SRSF1 and SRSF2 by ion-exchange chromatography and analyzing the fractions by SDS-PAGE. Fig. 2C and 2D show that both SR proteins are present in the *β-globin*:U1 snRNP complex that is eluted with ~ 500 mM NaCl (~ 14% buffer B); the unbound material was mostly aggregate and the last peak eluted contained only SR protein-bound RNA. This clearly suggests that SRSF1 plays two roles – modulation of pre-mRNA structure and recruitment of U1 snRNP to the pre-mRNA. Other SR proteins tested here can serve to modulate the pre-mRNA structure but cannot recruit U1 snRNP.

Multifaceted interactions between U1 snRNP and SRSF1 are keys to U1 snRNP recruitment

We previously reported that interaction between SRSF1-RBD and SNRP70-RBD, a component protein of U1 snRNP, is essential for assembly of E-complex (16). We also reported that I32 and V35 of SRSF1-RBD are essential for this interaction. We now investigate the interaction between U1 snRNP and SRSF1 with additional mutants of SRSF1-RBD, which are splicing defective. We previously reported that two acidic patch mutants of SRSF1-RBD (E62A/D63A/D66A i.e. EDD mutant, and E68A/D69A i.e. ED mutant) are not defective in ESE binding but defective in splicing (21). Therefore, we examined if these mutants are defective in U1 snRNP binding by pull-down assay. We also used F56D/F58D, a mutant of SRSF1 defective in RNA binding, and splicing (16, 22, 23). Fig. 3A shows that unlike WT SRSF1-RBD, neither mutant was able to pull-down U1 snRNP (compare lanes 13, 19, 20, 21). Interestingly, the groups of residues mutated in individual protein variants are significantly far from each other as displayed on the SRSF1-RRM1 solution structure (PDB code 1X4A) (Fig. 3B). This indicates that SRSF1 uses multiple surfaces to interact with U1 snRNP. Fig. 3A also shows that each of the protein variants can successfully block the pull-down of U1 snRNP by WT SRSF1-RBD (compare lanes 7, 8, 9, 13), conforming to the conclusion that SRSF1 uses multiple surfaces to interact with U1 snRNP.

We then examined the ability of these SRSF1-RBD variants to bind full-length *β-globin* and recruit U1 snRNP. Fig. 3C shows that both acidic patch mutants exhibit less cooperative binding to the full-length *β-globin* than the WT SRSF1-RBD (compare lanes 2, 4, 6). As expected F56D/F58D did not bind the pre-mRNA (lane 8). Both EDD and ED mutants could recruit U1 snRNP albeit less efficiently than the WT SRSF1-RBD (compare lanes 3, 5, 7). As expected, F56D/F58D did not recruit U1 snRNP at all (compare lanes 3 and 9). Overall, these results suggest that SRSF1 uses distinct but overlapping surfaces for interacting with pre-mRNA, U1 snRNP, and itself, and these interactions are essential for stabilization of U1 snRNP on the pre-mRNA.

5′SS-independent contacts are sufficient to initially recruit U1 snRNP

Next, we examined if the structural modulation of the pre-mRNA by SR proteins allows U1 snRNP to bind pre-mRNA specifically at the 5′SS. In the accompanying manuscript, we showed that the *β-globin* EH3+4 mutant is splicing defective and does not undergo extensive structural modulation in the presence of SR proteins unlike the WT substrate. In presence of all combinations of SR proteins tested here, U1 snRNP bound to the full-length *β-globin* as well as its mutant variants, Δ5′SS and EH3+4 (Fig. 4A). Δ5′SS variant of *β-globin* contained mutated authentic 5′SS and mutated cryptic 5′SS at -38, -16 and +13 position (24); the original and mutated sequences are shown in Fig. 5B. We also examined the specificity of U1 snRNP binding to *β-globin* by anion-exchange chromatography. For this study, we assembled the complex on WT and Δ5′SS *β-globin* with U1 snRNP and SRSF1-RE. Fig. 4B shows the chromatograms of WT complexes (blue line) and Δ5′SS complexes (red line). SDS-PAGE analysis indicates that peaks 1 and 4 (the flowthrough) consist of aggregates, peaks 2 and 5 majority of the ternary complex, and peaks 3 and 6 mostly free RNA (Fig. 4C). The complexes with both substrates were assembled with the same concentrations of pre-mRNAs and other components. Therefore, the similar extent of Coomassie staining of the protein components on either substrate indicates that similar constituents participate in either complex. This negates the possibility of the WT substrate binding one

5′SS-specific and one 5′SS-independent U1 snRNP and the mutant RNA binding only the 5′SS-independent U1 snRNP.

We then examined if deletion of the 3′ exon makes it possible to detect 5′SS-specific U1 snRNP binding with *βg-ΔEx2* constructs (i.e. *β-globin* with 3′ exon deleted). EMSA showed that *βg-ΔEx2* does not bind U1 snRNP specifically in the absence of SRSF1 (Fig. 4D, lane 5 and 6); SRSF1-dependent U1 snRNP binding to *βg-ΔEx2* is also dependent on presence of both authentic 5′SS and the exonic unpaired elements (Fig. 4D, compare lanes 3 & 4, 9 & 10, and 15 & 16). These data indicate that the SR proteins-mediated pre-mRNA structural modulation creates the context for U1 snRNP binding, which involves elements of both exons. The binding of U1 snRNP to *βg-ΔEx2* appeared to be weaker than that to full-length *β-globin*. We next assembled the complex with unlabeled *βg-ΔEx2*, SRSF1-RE, and U1 snRNP and attempted to purify it by anion-exchange chromatography. Fig. 4E shows the chromatograms of purification of *βg-ΔEx2* (blue line) and *βg-ΔEx2* Δ5′SS (red line) complexes, which were identical for both substrates. Remarkably, we detected protein components only in the aggregates present in the flowthrough fractions but none of the elution-peaks with both substrates (Fig. 4F). This strongly suggests that stable U1 snRNP binding to the pre-mRNA requires both exons. The presence of two peaks (3 & 4; 5 & 6) in the chromatograms without any protein components could be due to conformation of the RNA in solution, induced by U1 snRNP in presence of SRSF1, although U1 snRNP binding was not specific/stable enough and formed the aggregation with the pre-mRNA and SRSF1 during chromatographic purification. To test if base-pairing between U1 snRNA and 5′SS is essential for stabilizing U1 snRNP on the full-length pre-mRNA at this stage, we digested the 5′ end of U1 snRNA of the assembled and purified U1 snRNP particle with DNA-directed RNase H digestion, and then examined the binding by EMSA. The results showed that removal of the 5′ end of U1 snRNA did not affect the binding of U1 snRNP to full-length substrate indicating little contribution of base-pairing for U1 snRNP binding at this stage (Fig. S3, compare lanes 4 & 8); this result is consistent with little dependence of initial U1 snRNP binding on the presence of 5′SS (Fig. 4A).

Therefore, we postulate that the protein components of U1 snRNP provide sufficient support to stabilize U1 snRNP on the pre-mRNA at this stage. Role of SNRPC has been demonstrated in stabilizing U1 snRNP at the 5′SS (4). Here we examined the contribution of SNRPA in stabilizing U1 snRNP on the pre-mRNA. U1 snRNP B_{174} assembled with just the RRM1 (residues 1-101) of SNRPA (U1 snRNP $B_{174}A_{101}$) formed only a weak complex compared to U1 snRNP B_{174} since discrete complex was difficult to observe (Fig. 4G, compare lanes 5 & 10). At 480 nM U1 snRNP $B_{174}A_{101}$, the ternary complex was completely dismantled releasing free pre-mRNA; at this concentration, U1 snRNP B_{174} remained bound to the pre-mRNA (Fig. 4G, compare lanes 7 & 12). This suggests that the C-terminal RRM of SNRPA contributes to binding of U1 snRNP in a base-pair-independent manner, which is essential for stabilizing U1 snRNP on the pre-mRNA at this stage. We then examined interaction between U1 snRNP and SRSF1 using U1 snRNP B_{174} or U1 snRNP $B_{174}A_{101}$ by anion-exchange chromatography and SDS-PAGE analysis of the peak fractions. We assembled a complex with either U1 snRNP variant and SRSF1-RBD and purified chromatographically. Fig. S3B and S3C show that while U1 snRNP B_{174} remained bound to SRSF1-RBD, U1 snRNP $B_{174}A_{101}$ did not. This data indicates that in addition to SNRP70 (16), SNRPA also participates in stabilizing the interaction between U1 snRNP and SRSF1.

Based on these results, we conclude the following: first, the exonic unpaired elements upstream of 5’SS promotes SRSF1-mediated U1 snRNP recruitment; second, contacts involving both the exons stabilize U1 snRNP on the pre-mRNA leading to formation of the *bona fide* complex; third, intricate network of interactions involving SRSF1, U1 snRNP, and pre-mRNA initially recruits U1 snRNP to the pre-mRNA, and 5′SS recognition steps involve more contacts than just U1 snRNA:5´SS base-pairing. This conforms to our previous observation that *β-globin* ED1+2 mutants with completely single-stranded authentic 5′SS, which can readily base-pair to the 5′ end of U1 snRNA, is not sufficient for assembly of the spliceosome without functional exonic unpaired elements (EH3+4+ED1+2 mutant of *β-globin*, the accompanying manuscript).

Assembly and characterization of the earliest known spliceosomal complex incorporating all four major splice signals

We next examined how specific binding of U1 snRNP is ensured before the entry of a substrate into the spliceosome, given that U1 snRNP can bind the pre-mRNA in a 5′SS-independent manner. We hypothesized that 3′SS would only be recognized only when U1 snRNP is bound at the authentic 5′SS. Therefore, we examined if U2AF65, U2AF35, and SF1 bind the pre-mRNA in a splice signal specific manner after recruitment of U1 snRNP. For pre-mRNAs, we initially used two variants of *β-globin*, WT and a null PPT mutant with all pyrimidines replaced with purines (ΔPPT). To the pre-mRNA, we added either full-length or truncated (UHM domain, 38-152 a.a.) U2AF35, $SF1_{1\text{-}320}$ (25), full-length U2AF65, SRSF1-RBD or SRSF1-RE, and a near native assembled and purified U1 snRNP variant ($SNRP70_{1\text{-}215}$). Some lanes additionally contained hnRNP A1 since a recent report suggests that hnRNP A1 proofreads 3′SS for U2AF heterodimer (6). Fig. S4A shows that in the absence of U1 snRNP, no combination of proteins could distinguish between WT and ΔPPT substrates (compare lanes 2 and 11, 4 and 13, 6 and 15, 8 and 17). However, in presence of U1 snRNP, different combinations of proteins showed varied level of specificity, with the combination of U2AF35 UHM (38-152 a.a.), U2AF65, $SF1_{1\text{-}320}$, hnRNP A1, and SRSF1-RBD showing the highest specificity for recognizing PPT (compare lanes 5 and 14). Next, we examined if inclusion of SRSF1-RE and full-length U2AF35 instead of SRSF1-RBD and U2AF35-UHM domain, respectively, improve the specificity for PPT. Fig. S4B shows that indeed full-length variant of both proteins improves the specificity for PPT (compare lanes 7 and 14). Therefore, even though U2AF35 (26) and the RS domain of SRSF1 (15) are dispensable for splicing of *β-globin* in nuclear extract, all domains of these proteins contribute to specific recognition of PPT.

We next examined if the combination of proteins indicated above can specifically recognize all known splice signals and if the exonic unpaired elements upstream of the 5′SS, which we have described in the accompanying manuscript, is required. Fig. 5A shows that this set of proteins specifically distinguishes between WT and mutant splice signals (compare lanes 2, 4, 6, & 8; 12 & 14). In contrast, these factors only modestly distinguish between WT and mutant 3′SS (Δ3′SS) (compare lanes 2 and 10). We have shown in the accompanying manuscript that *β-globin* EH3+4 mutant has abolished splicing *in vivo*. Here we have examined the splicing efficiency of the other four mutants. Fig. 5B shows that Δ5′SS and ΔPPT mutants of *β-globin* do not splice, Δ3′SS show diminished splicing, and ΔBS splices efficiently. *β-globin* is known to have a cryptic BS, which is activated upon mutagenesis of the authentic BS (27). We surmise that splicing factors tested in Fig. 5A do not recognize the cryptic BS very well; recognition of the cryptic BS might require additional factors. On the other hand, Sanger sequencing of the spliced mRNA of Δ3′SS *β-globin* showed that the mRNA was synthesized from a cryptic 3′SS (AG) 26-nt downstream of the authentic 3′SS. We surmise that in this case, the splicing factors tested here recognized the downstream cryptic 3′SS in the absence of the authentic 3′SS. Next, we examined if SRSF2 lowers the level of SRSF1 required for 3′SS recognition. Fig. 5C shows that indeed SRSF2 lowers the level of SRSF1 required for 3′SS recognition. Interestingly, the complex assembled on ΔBS in presence of SRSF2 is stronger than the ones formed on Δ5′SS or ΔPPT (compare lanes 6 with 4 & 8). This conforms to our hypothesis that recognition of the cryptic BS in ΔBS *β-globin* requires additional factors. We named this complex with all major splice signals recognized as the recognition (R) complex.

Next, we examined if the base-pairing between the 5′SS and U1 snRNP is essential for assembly of R-complex. We digested the 5′ end of U1 snRNA of assembled and purified near-native U1 snRNP by DNA-directed RNase H digestion and then added the 3′SS recognition factors to WT *β-globin*. Fig. 5D shows that occlusion of the 5′ end of U1 snRNA with complementary DNA partly interrupts R-complex assembly; RNase H mediated digestion of the 5′ end of U1 snRNA almost completely abolished R-complex assembly, producing only U1 snRNP-recruited complex as shown before (Fig. 1A, S3A). These results indicate that 5′SS is incorporated into the R-complex through base-pairing of U1 snRNA to the 5′SS.

To further understand the specificity of the R-complex, we purified the complexes formed on WT *β-globin* and ΔPPT mutant by anion-exchange chromatography and analyzed the peak fractions by SDS-PAGE. Fig. 5E shows the chromatograms of purification of both complexes, which could separate the non-specific aggregates formed on both substrates from the specific complex formed only on the WT pre-mRNA. SDS-PAGE analysis of the peaks showed presence of U1 snRNP, SRSF1-RE, $SF1_{1-320}$, U2AF65, and U2AF35 in the specific complex formed on the WT substrate as well as aggregates formed on the both substrates (Fig. 5F). SNRP70 of the aggregated sample did not resolve properly in the SDS gel. We also assembled R-complex on *AdML*, purified by anion-exchange chromatography, and analyzed by SDS-PAGE (Fig. S4C, D).

These results suggest that the four major authentic splice signals (5′SS, BS, PPT, and 3′SS) are recognized at the earliest stage of spliceosome assembly with the help of the 5′SS proximal exonic unpaired elements that mediate functional interactions between pre-mRNAs and SR proteins. Splice signal-specific recruitment of these factors to the pre-mRNA is essential for assembly of E-complex. Since E-complex assembly has been carried out with crude nuclear extracts so far, where all splicing components are present, most likely recruitment could not be stalled at the R-complex stage before the complex progressed to the E-complex.

DISCUSSION

Metazoan *cis*-acting splice signals that define an intron during spliceosome assembly contain highly degenerate nucleotide sequences and are defined both by sequence and context within the pre-mRNA. Lack of knowledge about the minimal constituents essential for SS recognition prevented a clear understanding of the interdependency between the sequence and the context. The present work reports *in vitro* stepwise reconstitution of a stable early spliceosomal complex that includes and requires the five major elements that defines a pre-mRNA – 5′SS, BS, PPT, 3′SS, and the newly identified exonic unpaired elements, rich in single stranded nucleotides (the accompanying manuscript). We observe that pre-mRNA-specific SR protein recruitment results in structural modulation of the pre-mRNA and subsequent U1 snRNP recruitment. Among all tested SR proteins, SRSF1 plays the dual role in U1 snRNP recruitment: it can modulate the pre-mRNA structure in an exonic unpaired elements-dependent manner and it is essential for stabilizing U1 snRNP on the pre-mRNA remodeled by another SR protein. Whether U1 snRNP stabilization on the remodeled pre-mRNA is an exclusive function of SRSF1 or other untested SR proteins also could do this needs further analysis. Based on our results, we hypothesize that SR-protein binding to the pre-mRNA, with some SRSF1 molecules bound to both exons at both ends of the intron, renders a remodeled pre-mRNA that could engage U1 snRNP. The U1 snRNP recruitment is facilitated by interactions of SRSF1 molecules with U1 snRNP, pre-mRNA, and itself (Fig. 6). Furthermore, we observe that in the absence of U1 snRNP, the 3′SS recognition factors (U2AF65, U2AF35, and SF1) do not exhibit specific binding to the pre-mRNA:SR protein complex. We hypothesize that U1 snRNP binding constrains the pre-mRNA in a state that likely displays the 3′SS for specific interactions. However, the high degeneracy of the 3′SS nucleotide sequence implies that its recognition is dependent on not only a properly displayed 3′SS sequence, but also proper positioning of the 3′SS recognition factors guided by factors other than the splice site nucleotide sequence. Therefore, we hypothesize that the 3′SS recognition factors specifically interact not only with the splice signals, but also with other protein components of the pre-mRNA:SR:U1 snRNP complex, which would possibly require the 5′SS and 3′SS to remain in close proximity. We named this early spliceosomal complex as the ‘recognition (R) complex’ (Fig. 6).

These observations raise various critical questions. We observe that constitutive pre-mRNA substrates require exonic unpaired elements-dependent pre-mRNA structural modulation for U1 snRNP recruitment where the complementarity between the 5′SS nucleotide sequence and the 5′ end of U1 snRNA is high. This indicates that binding of U1 snRNP to the 5′SS can occur only under an appropriate structural context as has been suggested before (28). We suggest that, SR protein-mediated structural modulation of the pre-mRNA provides the context of the 5′SS. U1 snRNP binding independent of 5′SS or the exonic unpaired elements upstream of 5′SS to full-length *β-globin* (Fig. 4A) indicates that elements outside of 5′ exon or 5′SS also participate in defining the context for U1 snRNP binding. A previous genome-wide bioinformatic analysis of about 350,000 5′SS indicates that 5′SS are highly structured genome-wide (29). Furthermore, in the accompanying manuscript we observed that a *β-globin* pre-mRNA with hybridized 5′ exon with its 5′SS on a single-stranded segment (EH3+4+ED1+2 mutant of *β-globin*) is still unable to assemble spliceosome even though the 5′SS resemble the consensus sequence and considered strong. Together, these could mean that the context of the 5′SS is defined by SR protein-mediated structural modulation before 5′SS nucleotide sequence is accessible to U1 snRNA for base-pairing. Regarding SR protein-mediated structural modulation, SRSF1, SRSF2, and SRSF5 demonstrate compensatory nature (20) in recruitment of U1 snRNP to constitutively spliced *β-globin in vitro* (i.e. an SR protein can be replaced by another when the former is depleted). This backup feature likely limits the loss of constitutive splicing when a factor is missing. In contrast, absence of a specific SR protein essential for a specific regulated splicing event could have a drastic effect on the latter.

Nonetheless, requirement of a strong spatial context for recognition of 5′SS raises the next question: what happens when the context of the 5′SS is initially defined within an early spliceosomal complex, but base-pairing potential of 5′SS is absent or reduced due to mutations. Our results indicate that null mutation of the 5′SS causes the spliceosomal R-complex to be arrested in an aggregate-like complex, upon interaction with the 3′SS recognition factors. We are not clear what entails this aggregation but previous studies of U1 snRNP binding to the spliceosome in the absence of 5′SS indicates similar interactions (30). Interestingly, similar aggregate formation is also observed when any of the other splice signals were mutated; this suggests that the ‘splicing factors’-mediated context is necessary in conjunction with the nucleotide sequences of splice sites for their function.

Current compositional analysis of E-complex is not sufficient in characterizing requirements and roles of components essential for splice site recognition. In this study, we demonstrate the functional interdependence of splicing factors and splice signals in splice site recognition through assembly of the R-complex *in vitro*. In addition, we show that 5′ and 3′ splice sites crosstalk with each other across the intron at a very early stage of splice site recognition. Thus, this work provides a basis for further mechanistic dissection of molecular mechanisms underlying splice site recognition and its errors leading to diseases.

ACKNOWLEDGEMENT

This work was supported by NIH grant GM 085490 to GG.

REFERENCES

1. Will CL & Luhrmann R (2011) Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* 3:pii: a003707.
2. Mount SM, Pettersson I, Hinterberger M, Karmas A, & Steitz JA (1983) The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro. *Cell* 33:509-518.
3. Zillmann M, Rose SD, & Berget SM (1987) U1 small nuclear ribonucleoproteins are required early during spliceosome assembly. *Mol Cell Biol* 7:2877-2883.
4. Kondo Y, Oubridge C, Roon AMv, & 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.
5. Roca X, Krainer AR, & Eperon IC (2013) Pick one, but be quick: 5′ splice sites and the problems of too many choices. *Genes Dev.* 27:129-144
6. Tavanez JP, Madl T, Kooshapur H, Sattler M, & Valcárcel J (2012) hnRNP A1 proofreads 3' splice site recognition by U2AF. *Mol Cell* 45:314-329.
7. Yoshida H*, et al.* (2015) A novel 3' splice site recognition by the two zinc fingers in the U2AF small subunit. *Genes Dev* 29:1649-1660.
8. Michaud S & 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.
9. Barabino SM, Blencowe BJ, Ryder U, Sproat BS, & Lamond AI (1990) Targeted snRNP depletion reveals an additional role for mammalian U1 snRNP in spliceosome assembly. *Cell* 63:293–302.
10. Kent OA & MacMillan AM (2002) Early organization of pre-mRNA during spliceosome assembly. *Nat Struct Biol* 9:576–581.
11. Makarov EM, Owen N, Bottrill A, & Makarova OV (2012) Functional mammalian spliceosomal complex E contains SMN complex proteins in addition to U1 and U2 snRNPs. *Nucleic Acids Res* 40:2639-2652.
12. Zhou Z & Fu X-D (2013) Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* 122:191-207.
13. Salz HK*, et al.* (2004) The Drosophila U1-70K protein is required for viability, but its arginine-rich domain is dispensable. *Genetics* 168:2059-2065.
14. Pomeranz-Krummel DA, Oubridge C, Leung AK, Li J, & Nagai K (2009) Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature* 458:475-480.

15. Shaw SD, Chakrabarti S, Ghosh G, & Krainer AR (2007) Deletion of the N-terminus of SF2/ASF Permits RS-Domain-Independent Pre-mRNA Splicing. *Plos One* 2:e854.
16. 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.
17. Kohtz JD*, et al.* (1994) Protein-protein interactions and 5′-splice-site recognition in mammalian mRNA precursors. *Nature* 368:119-124.
18. Eperon IC*, et al.* (2000) Selection of alternative 5′ splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol Cell Biol* 20.
19. 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.
20. Pandit S*, et al.* (2013) Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol. Cell* 50:223-235.
21. Cho S*, et al.* (2011) The SRSF1 linker induces semi-conservative ESE binding by cooperating with the RRMs. *Nucl. Acids Res* 39: 9413-9421.
22. Cáceres JF & Krainer AR (1992) Functional analysis of pre-mRNA splicing factor SF2/ASF structural domains. *EMBO J* 12:4715–4726.
23. Zuo P & Manley JL (1993) Functional domains of the human splicing factor ASF/SF2. *EMBO J* 12:4727-4737.
24. Roca X, Sachidanandam R, & Krainer AR (2003) Intrinsic differences between authentic and cryptic 5′ splice sites. *Nucl. Acids Res* 31:6321-6333.
25. Guth S & 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.
26. Guth S, Tange TØ, Kellenberger E, & Valcárcel J (2001) Dual function for U2AF35 in AG-dependent pre-mRNA splicing. *Mol Cell Biol* 21:7673-7681.
27. Ruskin B, Greene JM, & Green MR (1985) Cryptic branch point activation allows accurate *in vitro* splicing of human β-globin intron mutants. *Cell* 41:833-844.
28. Nelson KK & Green MR (1988) Splice site selection and ribonucleoprotein complex assembly during in vitro pre-mRNA splicing. *Genes Dev* 2:319-329.
29. Kawaguchi R & Kiryu H (2016) Parallel computation of genome-scale RNA secondary structure to detect structural constraints on human genome. *BMC Bioinformatics* 17:203.
30. Fu X-D & Maniatis T (1992) The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site. *Proc. Natl. Acad. Sci. USA* 89:1725-1729.
31. 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.
32. Zuo P & Maniatis T (1996) The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes Dev* 10:1356-1368.
33. Valcarcel J, Gaur RK, Singh R, & Green MR (1996) Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA. *Science* 273:1706-1709.
34. Chen L*, et al.* (2016) Stoichiometries of U2AF35, U2AF65 and U2 snRNP reveal new early spliceosome assembly pathways. *Nucleic Acids Res*:pii: gkw860.

Methods

Cloning and protein expression

cDNAs for different subunits of U1 snRNP and SR proteins were cloned from *HeLa* cells, purchased from Addgene or Open Biosystems or obtained as gifts from other laboratories. The cDNAs of all proteins except for that of full-length U2AF35 (U2AF35) were cloned into T7 promoter-based *E. coli* expression vectors and were expressed as non-fusion or hexa-histidine/glutathione-S-transferase fusion proteins. Fusion proteins contain a *TEV* protease cleavage site for removal of the tag. Tag-removed proteins were used in all experiments except for pull-down assays. 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, and purified by Ni-NTA (Ni^{2+}-nitrilotriacetate) or Glutathione sepharose (GE Healthcare Lifesciences) affinity chromatography. The GST- or His_6-tags were commonly 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). Sm core proteins were co-expressed in combinations (D3-B, D1-D2, and E-F-G) and purified with similar procedures as described before(4). U2AF65 and $SF1_{1-320}$ were co-expressed in *E. coli* and purified as a heterodimer. Four functional variants of SR proteins were used: RNA binding domain of SRSF1 (1-203 a.a.) (16), full-length hyperphosphorylated mimetic SRSF1 with all serines of the RS domain (197-246 a.a.) replaced with glutamate (16), RNA binding domain of SRSF5 (1-184 a.a.) , and RNA binding domain of SRSF2 (1-127 a.a.) in chimera with fully phosphorylated mimetic serine-arginine domain SRSF1 (197-246 a.a.) with all serines replaced with glutamate (16, 31). Full-length U2AF35 was expressed in baculovirus-infected Sf9 cells and purified under denaturing conditions as described before (32). All purified proteins were confirmed to be RNase-free by incubating a small aliquot of the purified protein with a long RNA (U1 snRNA or *β-globin* pre-mRNA) overnight at room temperature and analyzing the RNA quality by urea PAGE following phenol extraction.

Electrophoretic mobility shift assay (EMSA)

Pre-mRNA was uniformly labeled with [*α-P^{32}*]*UTP* (3000 Ci/mmol; 10 µCi/µl) using run off transcription driven by T7 RNA polymerase (New England Biolabs), treated with 2 units of DNase I (New England Biolabs) for 1 hr at 37^{o}C, desalted twice by Illustra Microspin G-25 columns (GE Healthcare Life Sciences), and stored in water at -20^{o}C. For EMSA, ~ 10 pM radiolabeled pre-mRNA was incubated with SR proteins for 20 min at 30 ^{o}C in 20 mM HEPES-NaOH (pH 7.5), 250 mM NaCl, 1 mM DTT, 2 mM $MgCl_2$, 1 M urea, 20% glycerol and 0.3 % polyvinyl alcohol (PVA) in 15 µl volume. After incubation with SR proteins, the probe was incubated with U1 snRNP and 3′SS recognition factors as indicated, for 5 min at 30 °C. Reaction products were resolved on 4 % (89:1) polyacrylamide gels containing 2.5 % glycerol and 50 mM Tris-glycine buffer. All gels were run at 250 V for 90 min at 4 ^{o}C, dried and analyzed by phosphorimaging.

Antibody super-shift was carried out with anti-SNRPC antibody (Abcam, ab157116) and anti-SRSF1 antibody (Life Technologies, 32-4500). After the formation of complexes as described above, 0.25 µg antibody was added to the reaction, incubated at 30 ^{o}C for 5 min and resolved on polyacrylamide gels.

For DNA-directed RNase H digestion of the 5′ end of the U1 snRNA in the assembled and purified U1 snRNP was carried out by incubating ~1.2 µM U1 snRNP with 850 nM DNA (5′AGGTAAGTA3′) complementary to the 5′ end of U1 snRNA at room temperature for 20 min with 1 unit of RNase H (New England Biolabs).

In vitro reconstitution and purification of RNA-protein complexes

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. U1 snRNP was assembled as described before (4) and purified by anion exchange chromatography (Mono Q; GE Healthcare Lifesciences) 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.

Pre-mRNA in complex with U1 snRNP, SRSF1, and 3′SS recognition factors were assembled under the EMSA conditions but in 2 ml scale. Briefly, 25 nM pre-mRNA was incubated with 200 nM SRSF1-RE, 120 nM U1 snRNP, 250 nM U2AF65, 250 nM SF1, 250 nM U2AF35, and 250 nM HNRNP A1. The complex was purified using Mono Q column using an NaCl gradient (250 mM NaCl through 2M NaCl).

For MBP-MS3 pull-down, complexes were assembled similarly on RNA with 3x MS3 binding sites at the 3′ end and was pulled down with MBP-tagged MS3-coat protein and amylose resin (New England Biolabs).

Electron microscopy

U1 snRNP was bound to Carbon-Formvar grid (01754-F F/C 400 mesh Cu from Ted Pella) activated by 30s of glow discharge, negatively stained with 0.5% Uranyl Acetate and imaged in FEI Tecnai G2 Sphera.

FIG. LEGENDS

Fig. 1. SRSF1-mediated U1 snRNP recruitment to *β-globin*: (A) Titration of SRSF1-RBD -saturated *β-globin* with U1 snRNP displaced SRSF1 molecules and formed U1 snRNP-dependent complexes (marked with arrows) (lanes 2-13); U1 snRNP did not complex with free pre-mRNA (lanes 14-16). (B) Titration of SRSF1-saturated *β-globin* with U1 snRNP B_{174} showing displacement of SRSF1 molecules and formation of U1 snRNP B_{174}-dependent complexes (marked with arrows) (lanes 2-4); U1 snRNP B_{174} does not complex with free pre-mRNA (lanes 5-6). (C) Super-shift of complexes with αSRSF1 and αSNRPC; comparison of lanes 14, 16 and 18 indicates that U1 snRNP-dependent complexes form at lower concentration of U1 snRNP than what is needed for exhaustive displacement of excess SRSF1 molecules bound to the pre-mRNA and resolution of U1 snRNP-dependent complexes on gel; U1 snRNP B_{174}-dependent complex super-shifted with αSNRPC (lanes 19 & 20). (D) Pull-down and SDS-

PAGE of MS3-tagged *β-globin* mixed with SRSF1-RE (lane 1), SRSF1-RE+U1 snRNP (lane 2), and U1 snRNP (lane 3) with MBP-tagged MS3 protein bound to amylose resin.

Fig. 2. Binding of SR proteins that primes the pre-mRNA for U1 snRNP recruitment shows compensatory nature. (A) SRSF2 and SRSF5, on their own, do not recruit U1 snRNP to *β-globin* unlike SRSF1 (compares lanes 5, 9, & 13). (B) 60 nM (lane 2) but not 20 nM (lane 3) SRSF1 can efficiently recruit U1 snRNP to *β-globin*; presence of additional SRSF2 and/or SRSF5 enable low concentration of SRSF1 to efficiently recruit U1 snRNP (compare lanes 8, 10, & 12). (C) Chromatogram of purification of the *β-globin* complex containing both SRSF1 and SRSF2 and U1 snRNP by anion-exchange chromatography. (D) SDS-PAGE analysis of the peak fractions shown in (C).

Fig. 3. SRSF1 has multifaceted interactions with U1 snRNP: (A) WT SRSF1-RBD but not mutant SRSF1-RBD variants (E62A/D63A/D66A or EDD, E68A/D69A or ED, F56D/F58D or FF-DD) can pull down U1 snRNP from solution (lanes 13-21); these mutant variants block effective pull-down by WT SRSF1-RBD (lanes 1-13). (B) Position of the mutated residues are shown on solution structure of SRSF1-RRM1 (PDB code 1X4A). (C) EMSA showing partial loss of ability of the EDD and the ED mutants and full loss of ability of the FF-DD mutant in U1 snRNP recruitment.

Fig. 4. Initial U1 snRNP recruitment requires multiple contacts involving both exons across the intron: (A) U1 snRNP is recruited to all three variants (WT, Δ5′SS, and EH3+4) of full-length *β-globin* by SRSF1 alone or in conjunction with SRSF2 and/or SRSF5 (complexes are marked with arrows). (B) Chromatogram of purification of the ternary complex consisting of SRSF1-RE and U1 snRNP assembled on WT *β-globin* and its Δ5′SS mutant; blue and red lines indicate WT and Δ5′SS complexes, respectively; peaks are numbered with color-coded digits matching the color of the lines. (C) SDS-PAGE of the peak fractions shown in (B). (D) Unmodified *βg-ΔEx2* but not *βg-ΔEx2* with 5′SS mutation

(*βg-ΔEx2 Δ5′SS*) nor *βg-ΔEx2* with EH3+4 mutation (*βg-ΔEx2 EH3+4*) formed U1 snRNP B_{174}-dependent complexes in presence of 60 nM SRSF1 (the complexes are marked with arrows and the position of the missing complexes in the defective RNAs are marked with curved brackets); U1 snRNP B_{174} shows no specific interaction with free RNA. (E) Chromatograms of purification of U1 snRNP:SRSF1:pre-mRNA complexes assembled with *βg-ΔEx2* (blue line) or *βg-ΔEx2 Δ5′SS* (red line) by anion-exchange chromatography. (F) SDS-PAGE of peak fractions shown in (E); peak numbers are color-coded matching the color of the lines. (G) U1 snRNP B_{174} but not U1 snRNP B_{174} A_{101} assembles stable and distinct complexes with *β-globin* in presence of SRSF1-RBD.

Fig. 5. Specific recognition of the 3′ end of the intron by cooperative action of U1 snRNP, SF1, U2AF65, U2AF35, and hnRNP A1: (A) Specific recognition of all known major splice signals leading to assembly of R-complex with *β-globin* (blue rectangle, lanes 1-2, 13-14), near abolition of complex formation in Δ5′SS (compare 2 & 4), ΔBS (lanes 2 & 6), and ΔPPT (lanes 2 & 8), EH3+4 i.e. hybridized exonic unpaired elements (lanes 12 & 14) mutants (red rectangle) with the best combination of splicing factors identified in S3B; Δ3′SS mutant showed significantly less defect in R-complex assembly (lanes 2 & 10). (B) Transfection-based splicing assay of WT, Δ5′SS, ΔBS, ΔPPT, and Δ3′SS mutants of *β-globin*; Δ5′SS and ΔPPT are completely splicing defective, ΔBS produces authentic mRNA spliced using a cryptic BS, and Δ3′SS splices from a cryptic 3′SS 26-nt downstream of authentic 3′SS; * indicates cryptic mRNA; original and mutated nucleotide sequences of authentic 5′SS, cryptic 5′SS at -38, -16 and +13-nt positions, BS, and 3′SS are shown at the bottom. (C) Comparison of specific recognition of all major splice signals in presence of both SRSF1 and SRSF2 in *β-globin*. (D) DNA-directed RNase H digestion of the 5′ end of near-native U1 snRNP (SNRP70 ΔRS) prohibits R-complex assembly (E) Ion-exchange chromatogram of purification of complex assembled with U1 snRNP, SRSF1-RE, $SF1_{1-320}$, U2AF65, U2AF35, and hnRNP A1 on WT or ΔPPT *β-globin*. (F) SDS-PAGE of peak fractions shown in (E).

Fig. 6: Proposed model for assembly of R-complex, an early spliceosome with both splice sites recognized: A full-length pre-mRNA with a hypothetical secondary structure and five essential elements for splicing (5′SS, BS, PPT, 3′SS, and exonic unpaired elements, abbreviated EUE) is shown. Recruitment of SRSF1 by EUE induces a structural modulation of the pre-mRNA that extends beyond the EUE region. U1 snRNP binds the RNA:protein complex involving both exons. This possibly brings the exons close to each other and constrains the pre-mRNA in a high-order state marking BS, PPT, and 3'SS for recognition by SF1, U2AF65, and U2AF35, respectively. The role of much of the intron, which could be highly variable in length, in recognition of splice sites across the intron is yet to be uncovered. N- and C-termini of U2AF65 are shown; the N-terminal RS domain of U2AF65 is known to contact the BS (33), while the C-terminal UHM domain interacts with SF1, bound at the BS. hnRNP A1 forms a complex at the 3′SS in presence of U2AF heterodimer, ensuring the specific binding of the heterodimer. Solid bidirectional arrows indicate the RNA-independent mutual interactions among SF1, U2AF65, and U2AF35. The broken bidirectional arrow indicates RNA-dependent interaction.

SUPPLEMENTARY FIGURE LEGENDS

Fig. S1. A schematic depicting our current understanding of splice site recognition. The pre-mRNA (represented with a hypothetical secondary structure) contains the exonic unpaired elements (EUE) (the accompanying manuscript) in addition to four major splice signals. 5′SS-specific association of U1 snRNP with the pre-mRNA (mediated primarily by base-pairing between 5′SS and U1 snRNA), and binding of SF1, U2AF65, and U2AF35 to the BS, PPT, and 3′SS, respectively, are essential for splice site recognition and assembly of a pre-E-complex (34). The N-terminal RS domain of U2AF65 interact with the BS and the C-terminal UHM domain interacts with SF1 bound to the BS. U2AF35-UHM domain interacts with U2AF65 and BS remains in close proximity of 5′SS. Solid bidirectional arrows indicate the RNA-independent mutual interactions among SF1, U2AF65, and U2AF35. HNRNP A1 has

been shown to complex with U2AF65 in an RNA-dependent manner at the 3′SS and proofread binding of U2AF65 to a PPT that is followed by AG dinucleotide. The broken bidirectional arrow indicates RNA-dependent interactions. SR proteins are expected to be present in the complex but their roles in the assembly remains unclear except for that SRSF1 helps stabilize U1 snRNP. Minimal components essential for splice site recognition and the exact composition and assembly pathway of the pre-E-complex are currently not known. The pre-E-complex acts as a substrate for U2 snRNP binding independent of BS, for assembly of E-complex. The functions of majority of the proteins identified in E-complex purified from the nuclear extract remains unclear.

Fig. S2. Characterization of U1 snRNP binding to the pre-mRNA. (A) SDS-PAGE of fractions U1 snRNP (left) and U1 snRNP B_{174} (right) eluted from Mono Q column at ~ 400 mM KCl; SDS PAGE analysis of unpurified U1 snRNP is shown as ‘input’. (B) (left) Negatively stained images of near-native U1 snRNP reconstituted using full-length protein components except for a truncated variant of SNRP70; the image indicates extremely high homogeneity of U1 snRNP particles; the EM grid was prepared by depositing particles on the negatively charged (for 30 sec) Carbon-Formvar grid (01754-F F/C 400 mesh Cu from Ted Pella), briefly washing twice with deionized water, and staining with 0.5% Uranyl Formate for 1 min. (right) A preliminary 2D-average analysis of 2527 particles into 48 classes using program EMAN2, indicates the distinctive Sm core ring and additional protuberances representing SNRPA, SNRP70 etc. (C) Recruitment of U1 snRNP B_{174} to *AdML* in presence of but not in absence of SRSF1; U1 snRNP-dependent complexes are indicated with arrows. (D) SRSF1-RBD and SRSF1-RE but not SRSF1-RERA or WT full-length SRSF1 can recruit U1 snRNP to *β-globin*.

Fig. S3: Initial U1 snRNP recruitment requires multiple contacts: (A) DNA-directed RNase H digestion of the 5′ end of U1 snRNA does not interrupt binding of U1 snRNP to *β-globin*; 14-nt long *β-globin* 5′SS (βg 5′SS, 5′GGGCAGGUUGGUAU3′) was premixed with U1 snRNP, where indicated,

before addition of U1 snRNP to the *β-globin*:SRSF1 complex. (B) Chromatograms of purification of the binary complex assembled with SRSF1-RBD and either U1 snRNP B_{174} (blue line) or U1 snRNP B_{174} A_{101} (red line) by anion-exchange chromatography. (C) SDS-PAGE analysis of the peak fractions shown in (B) shows strong binding of SRSF1 to U1 snRNP B_{174} but not U1 snRNP B_{174} A_{101}.

Fig. S4: Identification of minimal components essential for recognition of splice sites: (A) Comparison of specific recognition of PPT with various combinations of splicing factors examined with *β-globin* WT and ΔPPT indicates that a combination of U1 snRNP, SRSF1-RBD, U2AF65, $SF1_{1\text{-}320}$, UHM domain of U2AF35, and HNRNP A1 recognized PPT most specifically (compare lanes 5 and 14); original and mutated nucleotide sequences of PPT are shown at the bottom. (B) Comparison of specific recognition of PPT with various combinations of splicing factors examined with *β-globin* WT and ΔPPT indicates that the combination of U1 snRNP, SRSF1-RE, U2AF65, $SF1_{1\text{-}320}$, full-length U2AF35, and HNRNP A1 recognized PPT most specifically (compare lanes 7 and 14). (C) Mono Q chromatogram of purification of R-complex assembled on *AdML*. (D) Analysis of the peak fractions of (C) by SDS-PAGE.

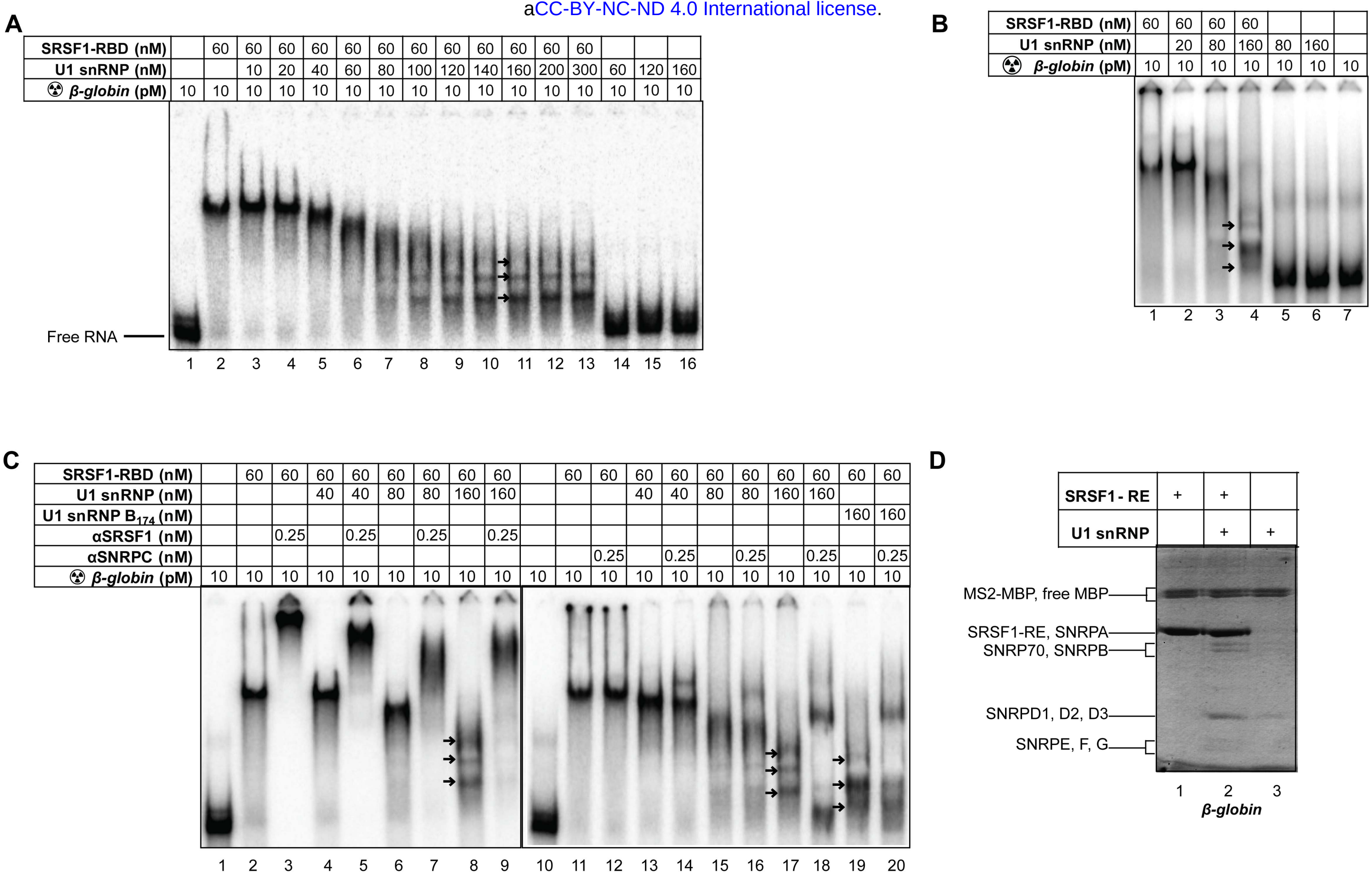
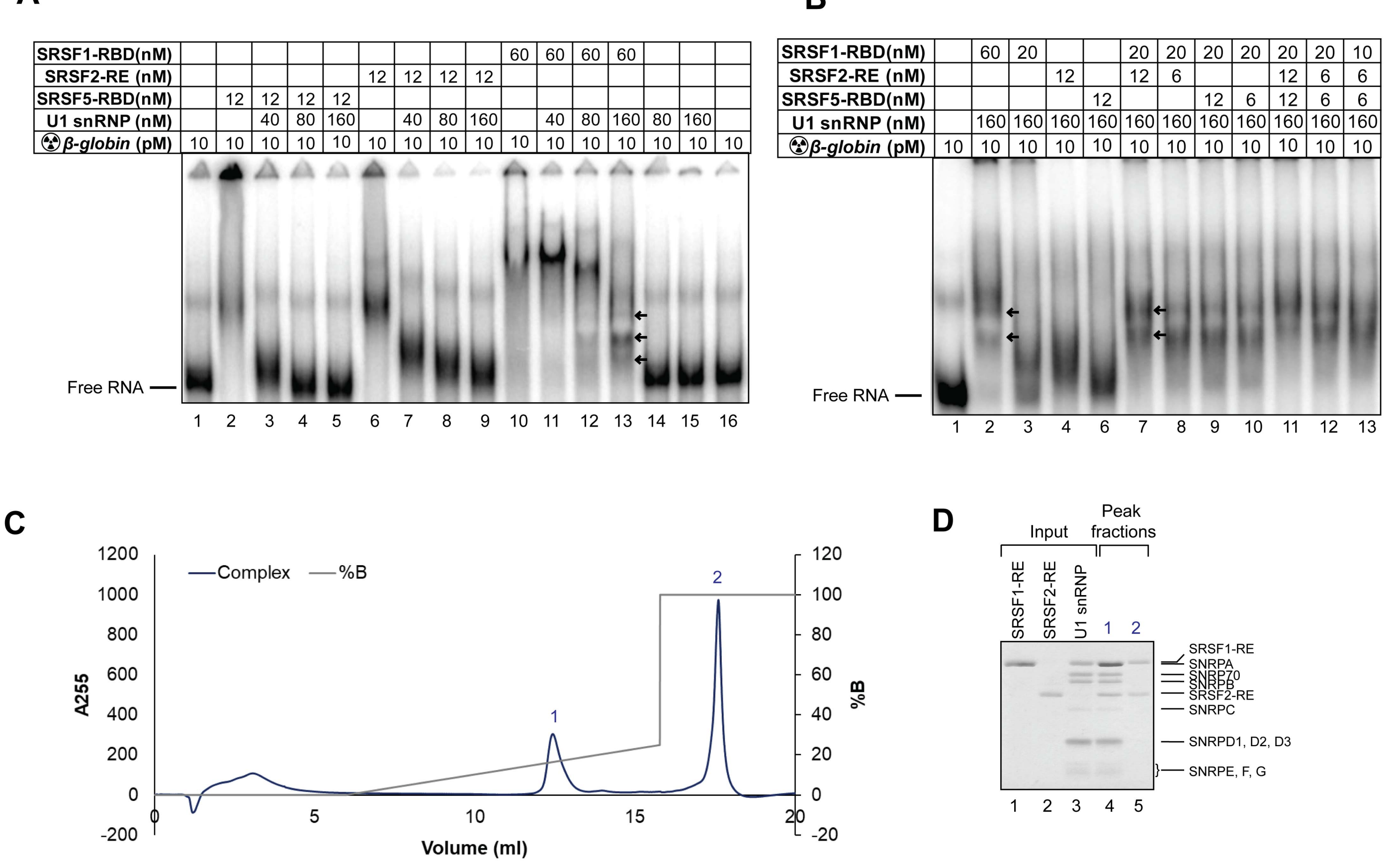

A
SRSF1-RBD (nM) | | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | | |
U1 snRNP (nM) | | | 10 | 20 | 40 | 60 | 80 | 100 | 120 | 140 | 160 | 200 | 300 | 60 | 120 | 160
β-globin (pM) | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10
Free RNA
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
B
SRSF1-RBD (nM) | 60 | 60 | 60 | 60 | | |
U1 snRNP (nM) | | 20 | 80 | 160 | 80 | 160 |
β-globin (pM) | 10 | 10 | 10 | 10 | 10 | 10 | 10
1 2 3 4 5 6 7
C
SRSF1-RBD (nM) | | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60 | 60
U1 snRNP (nM) | | | | 40 | 40 | 80 | 80 | 160 | 160 | | | | 40 | 40 | 80 | 80 | 160 | 160 | |
U1 snRNP B174 (nM) | | | | | | | | | | | | | | | | | | | 160 | 160
αSRSF1 (nM) | | | 0.25 | | 0.25 | | 0.25 | | 0.25 | | | | | | | | | | |
αSNRPC (nM) | | | | | | | | | | | | 0.25 | | 0.25 | | 0.25 | | 0.25 | | 0.25
β-globin (pM) | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
D
SRSF1- RE | + | + |
U1 snRNP | | + | +
MS2-MBP, free MBP
SRSF1-RE, SNRPA
SNRP70, SNRPB
SNRPD1, D2, D3
SNRPE, F, G
1 2 3
β-globin

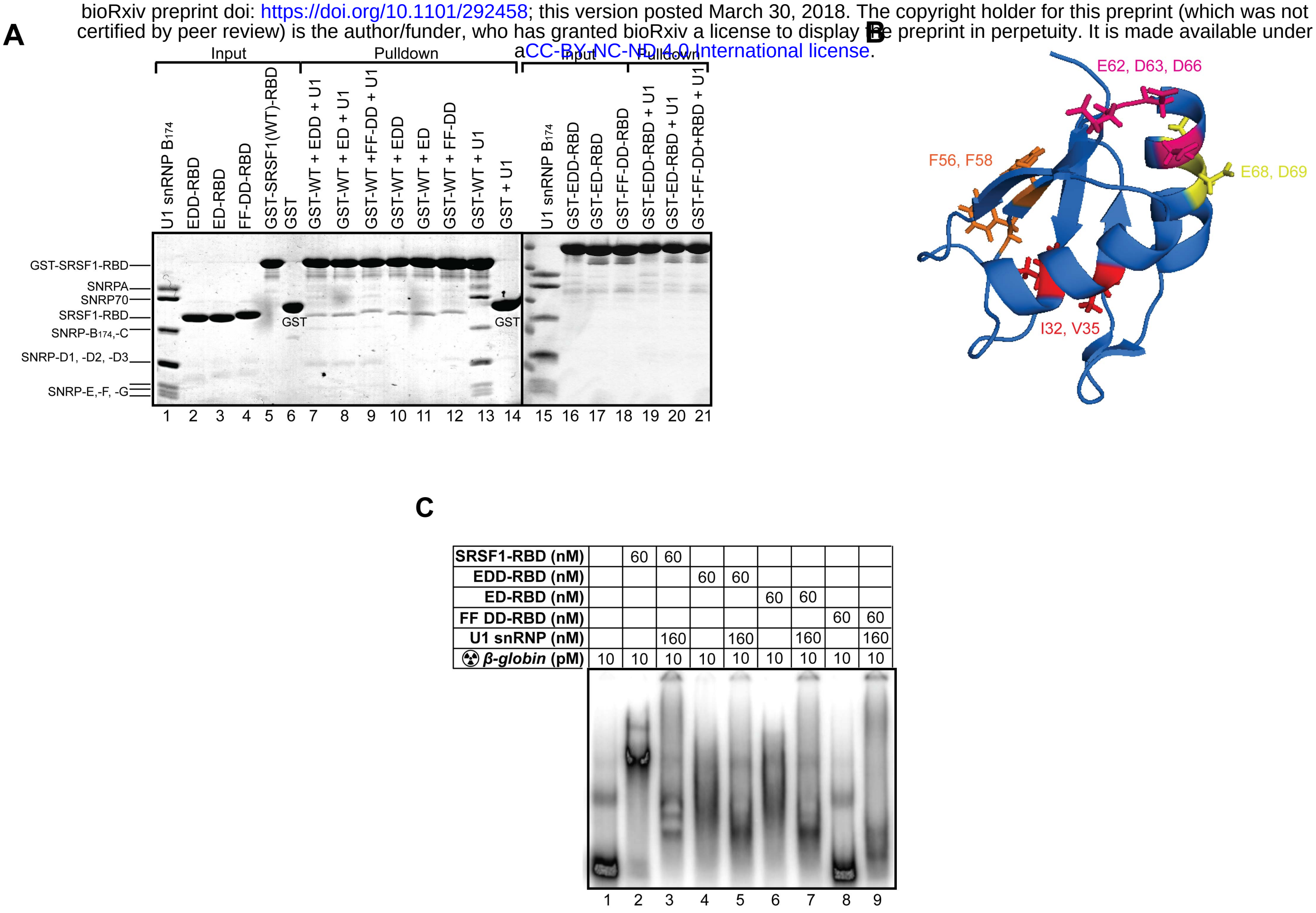
Figure 1



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
SRSF1-RBD(nM)										60	60	60	60			
SRSF2-RE (nM)						12	12	12	12							
SRSF5-RBD(nM)		12	12	12	12											
U1 snRNP (nM)			40	80	160		40	80	160		40	80	160	80	160	
☢ *β-globin* (pM)	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

	1	2	3	4	6	7	8	9	10	11	12	13
SRSF1-RBD(nM)		60	20			20	20	20	20	20	20	10
SRSF2-RE (nM)				12		12	6			12	6	6
SRSF5-RBD(nM)					12			12	6	12	6	6
U1 snRNP (nM)		160	160	160	160	160	160	160	160	160	160	160
☢ *β-globin* (pM)	10	10	10	10	10	10	10	10	10	10	10	10

Figure 2



SRSF1-RBD (nM)		60	60						
EDD-RBD (nM)				60	60				
ED-RBD (nM)						60	60		
FF DD-RBD (nM)								60	60
U1 snRNP (nM)			160		160		160		160
☢ *β-globin* (pM)	10	10	10	10	10	10	10	10	10

Figure 3

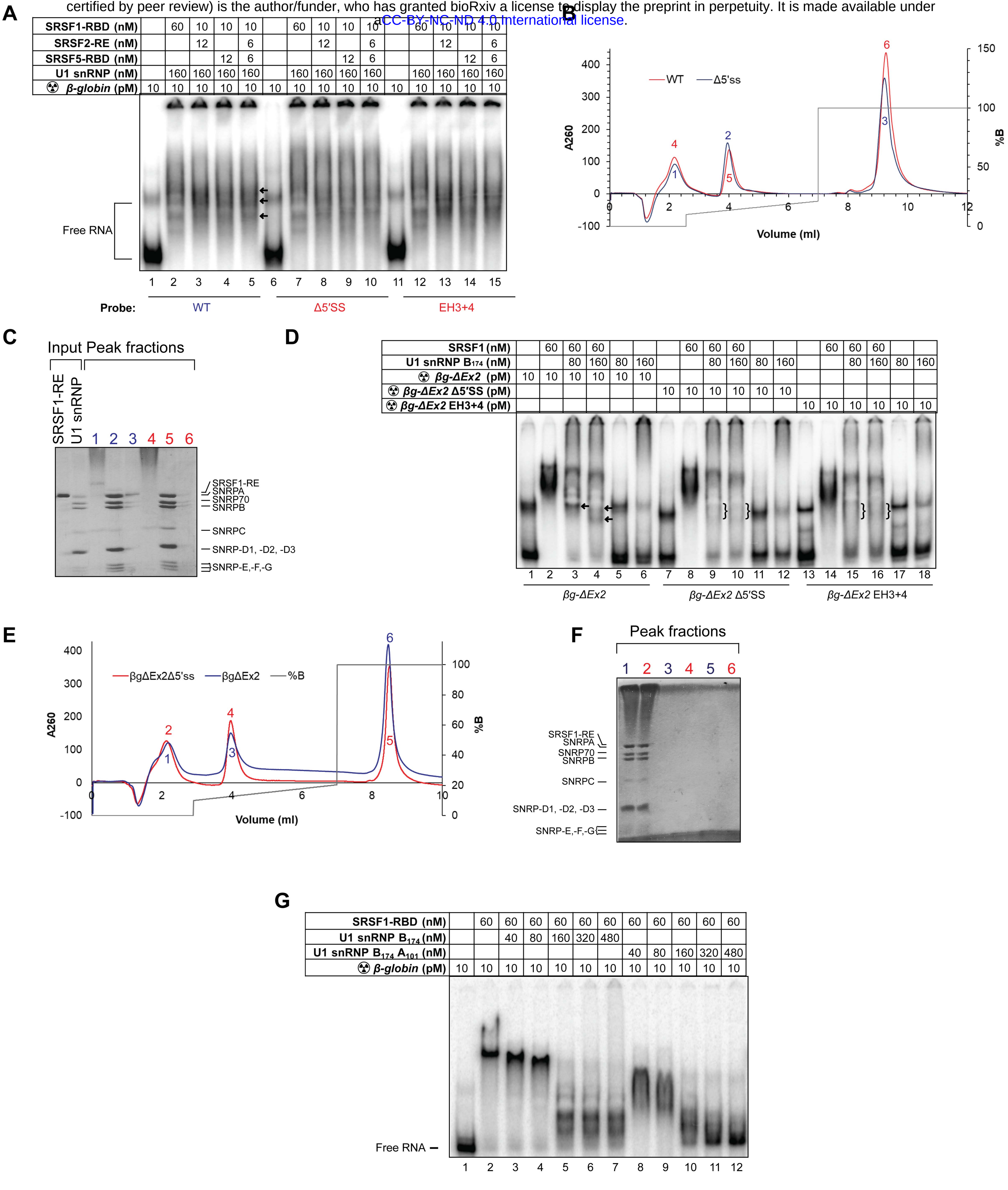
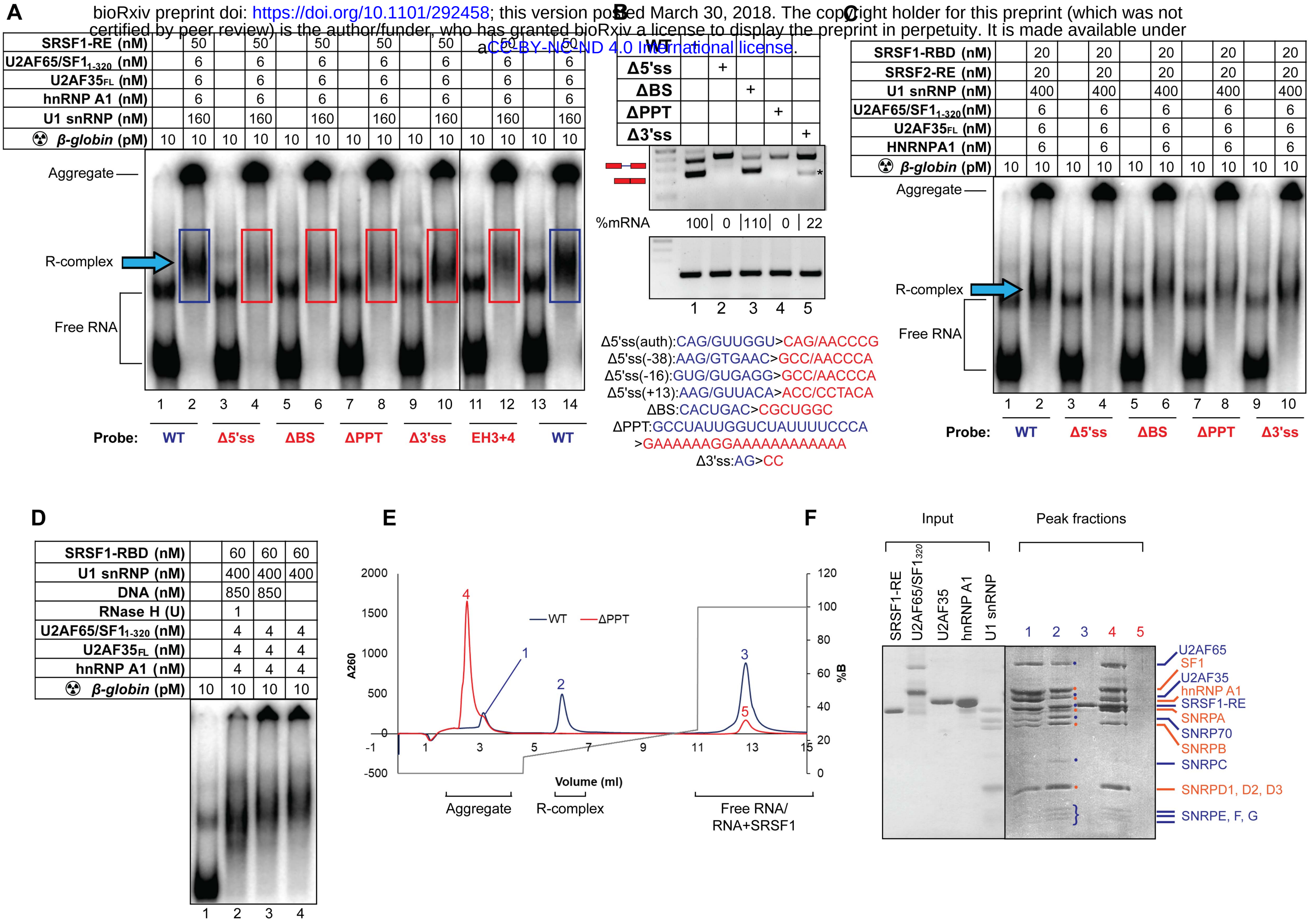

A
SRSF1-RBD (nM) 60 10 10 10 60 10 10 10 60 10 10 10
SRSF2-RE (nM) 12 6 12 6 12 6
SRSF5-RBD (nM) 12 6 12 6 12 6
U1 snRNP (nM) 160 160 160 160 160 160 160 160 160 160 160 160
β-globin (pM) 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
Free RNA
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Probe:
WT
Δ5′SS
EH3+4
B
WT
Δ5′ss
A260
%B
Volume (ml)
400
300
200
100
0
-100
150
100
50
0
0 2 4 6 8 10 12
1 2 3 4 5 6
C
Input Peak fractions
SRSF1-RE
U1 snRNP
1 2 3 4 5 6
SRSF1-RE
SNRPA
SNRP70
SNRPB
SNRPC
SNRP-D1, -D2, -D3
SNRP-E,-F,-G
D
SRSF1 (nM) 60 60 60 60 60 60 60 60 60
U1 snRNP B174 (nM) 80 160 80 160 80 160 80 160 80 160 80 160
βg-ΔEx2 (pM) 10 10 10 10 10 10
βg-ΔEx2 Δ5′SS (pM) 10 10 10 10 10 10
βg-ΔEx2 EH3+4 (pM) 10 10 10 10 10 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
βg-ΔEx2
βg-ΔEx2 Δ5′SS
βg-ΔEx2 EH3+4
E
βgΔEx2Δ5′ss
βgΔEx2
%B
A260
%B
Volume (ml)
400
300
200
100
0
-100
100
80
60
40
20
0
0 2 4 6 8 10
1 2 3 4 5 6
F
Peak fractions
1 2 3 4 5 6
SRSF1-RE
SNRPA
SNRP70
SNRPB
SNRPC
SNRP-D1, -D2, -D3
SNRP-E,-F,-G
G
SRSF1-RBD (nM) 60 60 60 60 60 60 60 60 60 60 60
U1 snRNP B174 (nM) 40 80 160 320 480
U1 snRNP B174 A101 (nM) 40 80 160 320 480
β-globin (pM) 10 10 10 10 10 10 10 10 10 10 10 10
Free RNA
1 2 3 4 5 6 7 8 9 10 11 12

Figure 4



SRSF1-RE (nM)		50		50		50		50		50		50		50
U2AF65/SF1$_{1-320}$ (nM)		6		6		6		6		6		6		6
U2AF35$_{FL}$ (nM)		6		6		6		6		6		6		6
hnRNP A1 (nM)		6		6		6		6		6		6		6
U1 snRNP (nM)		160		160		160		160		160		160		160
☢ *β-globin* (pM)	10	10	10	10	10	10	10	10	10	10	10	10	10	10

WT	+				
Δ5'ss		+			
ΔBS			+		
ΔPPT				+	
Δ3'ss					+

Δ5'ss(auth):CAG/GUUGGU>CAG/AACCCG
Δ5'ss(-38):AAG/GTGAAC>GCC/AACCCA
Δ5'ss(-16):GUG/GUGAGG>GCC/AACCCA
Δ5'ss(+13):AAG/GUUACA>ACC/CCTACA
ΔBS:CACUGAC>CGCUGGC
ΔPPT:GCCUAUUGGUCUAUUUUCCCA
>GAAAAAAGGAAAAAAAAAAAAA
Δ3'ss:AG>CC

SRSF1-RBD (nM)		20		20		20		20		20
SRSF2-RE (nM)		20		20		20		20		20
U1 snRNP (nM)		400		400		400		400		400
U2AF65/SF1$_{1-320}$(nM)		6		6		6		6		6
U2AF35$_{FL}$ (nM)		6		6		6		6		6
HNRNPA1 (nM)		6		6		6		6		6
☢ *β-globin* (pM)	10	10	10	10	10	10	10	10	10	10

SRSF1-RBD (nM)		60	60	60
U1 snRNP (nM)		400	400	400
DNA (nM)		850	850	
RNase H (U)		1		
U2AF65/SF1$_{1-320}$ (nM)		4	4	4
U2AF35$_{FL}$ (nM)		4	4	4
hnRNP A1 (nM)		4	4	4
☢ *β-globin* (pM)	10	10	10	10

Figure 5

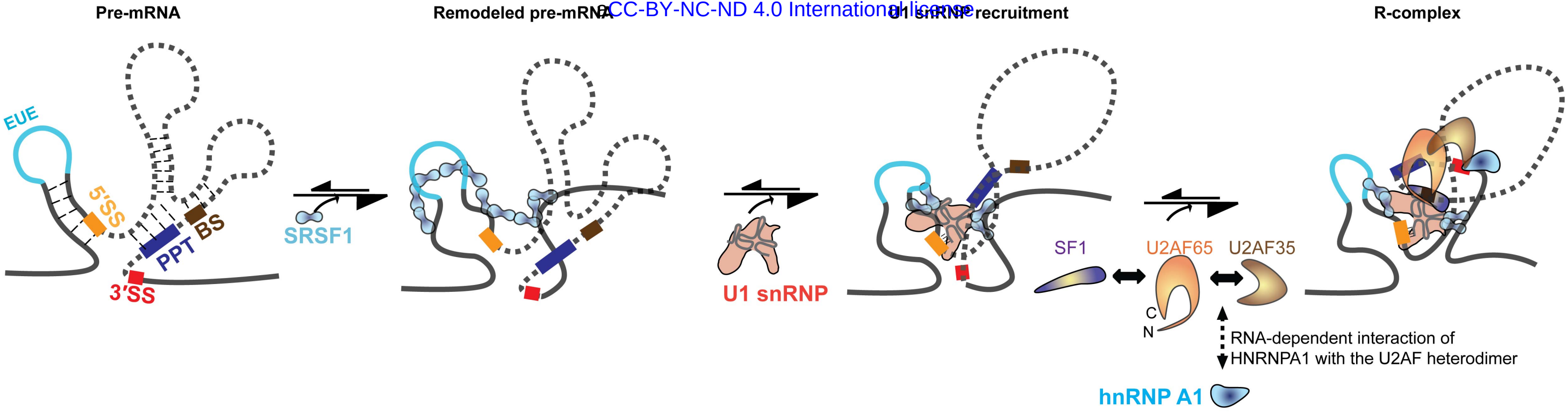

Pre-mRNA
Remodeled pre-mRNA
U1 snRNP recruitment
R-complex
EUE
5'SS
PPT
BS
3'SS
SRSF1
U1 snRNP
SF1
U2AF65
U2AF35
C
N
RNA-dependent interaction of
HNRNPA1 with the U2AF heterodimer
hnRNP A1

Figure 6

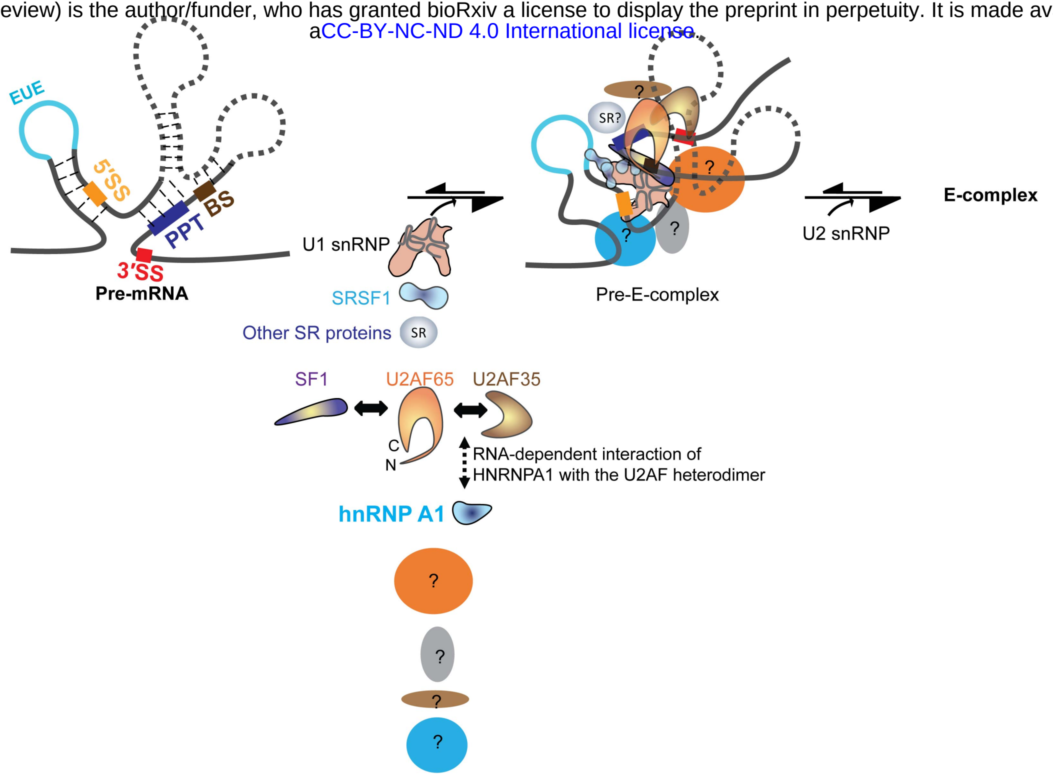
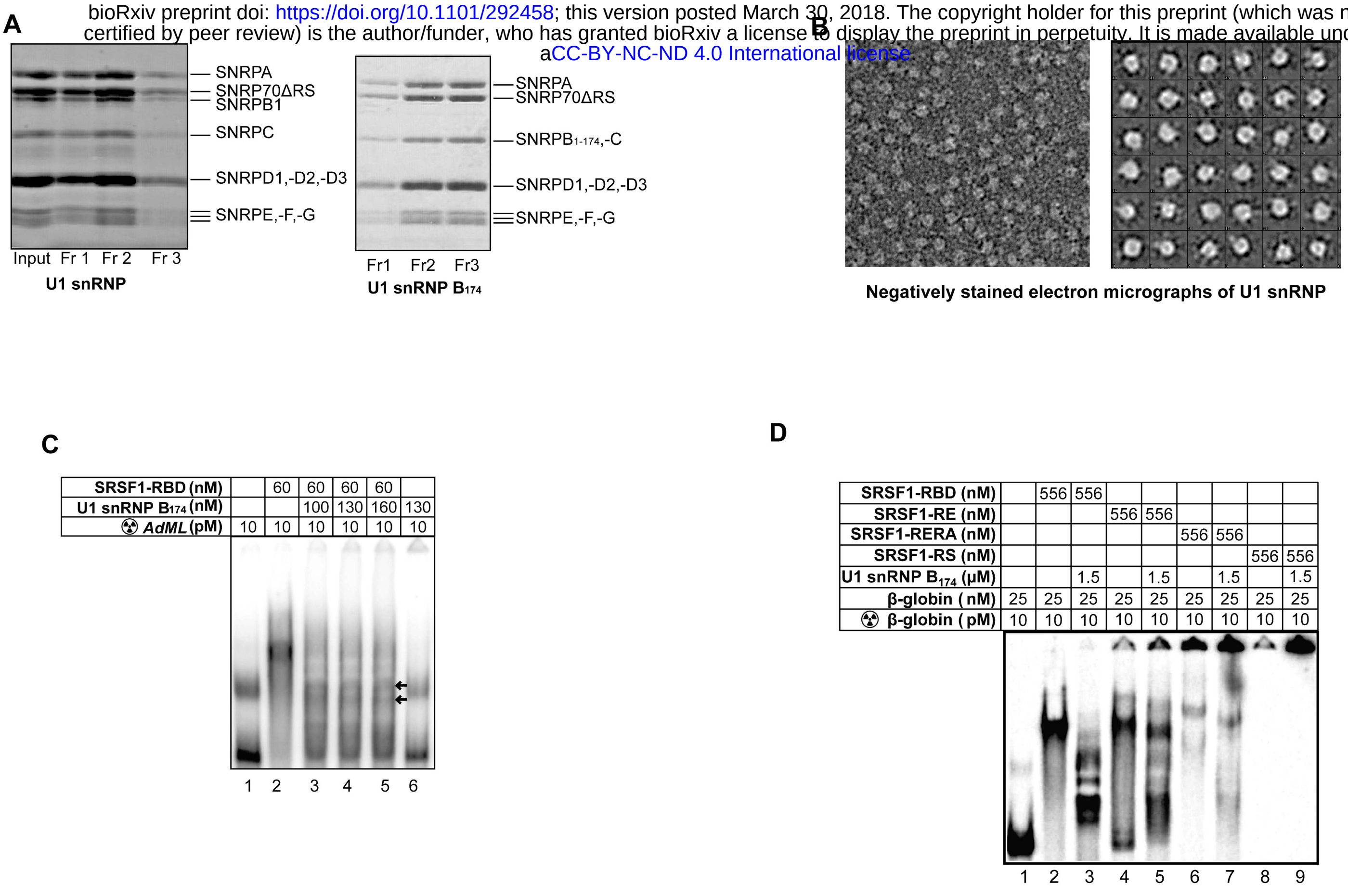

EUE
5'SS
PPT
BS
3'SS
Pre-mRNA
U1 snRNP
SRSF1
Other SR proteins
SR
SR?
?
Pre-E-complex
U2 snRNP
E-complex
SF1
U2AF65
U2AF35
C
N
RNA-dependent interaction of
HNRNPA1 with the U2AF heterodimer
hnRNP A1
?
?
?
?

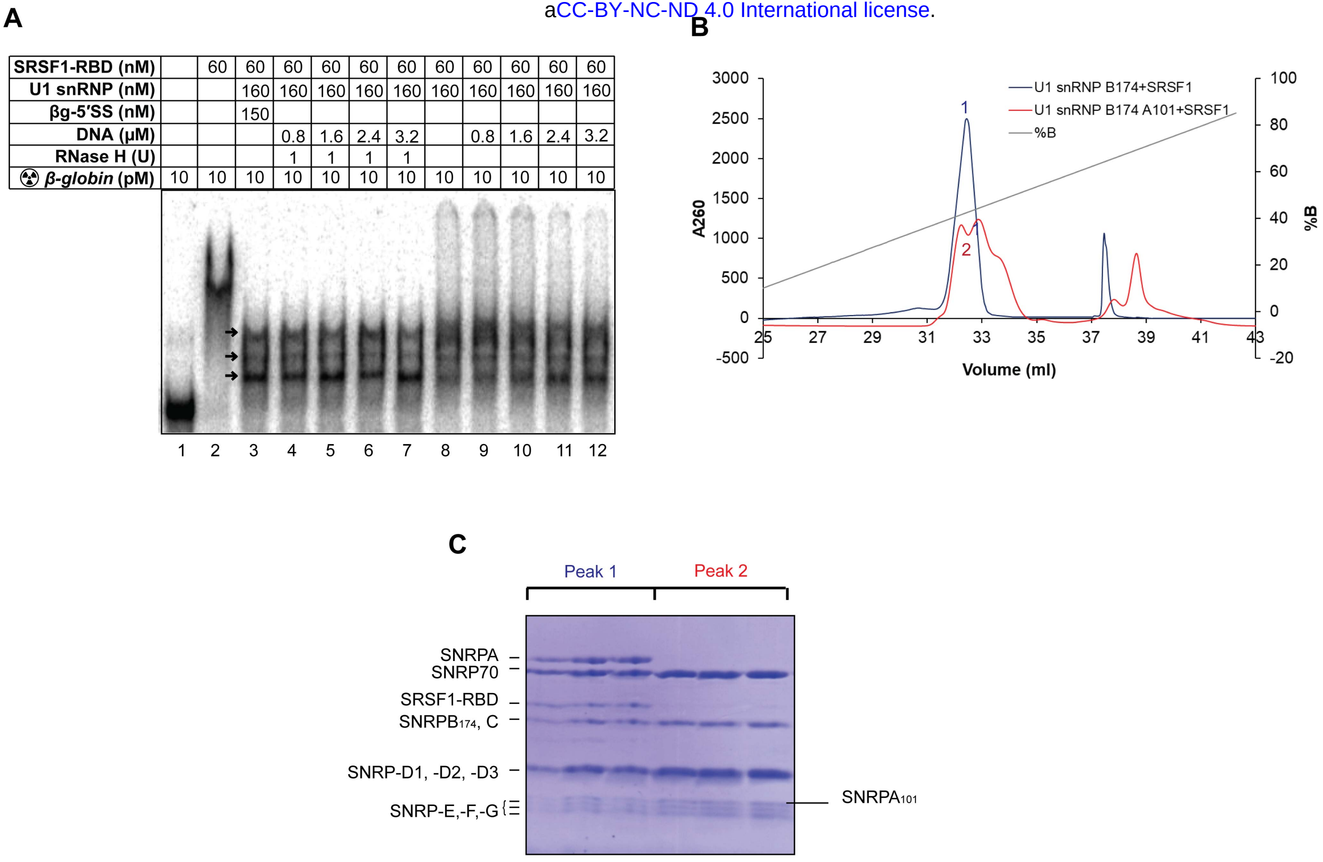
Figure S1



SRSF1-RBD (nM)		60	60	60	60	
U1 snRNP B174 (nM)			100	130	160	130
☢ ***AdML* (pM)**	10	10	10	10	10	10

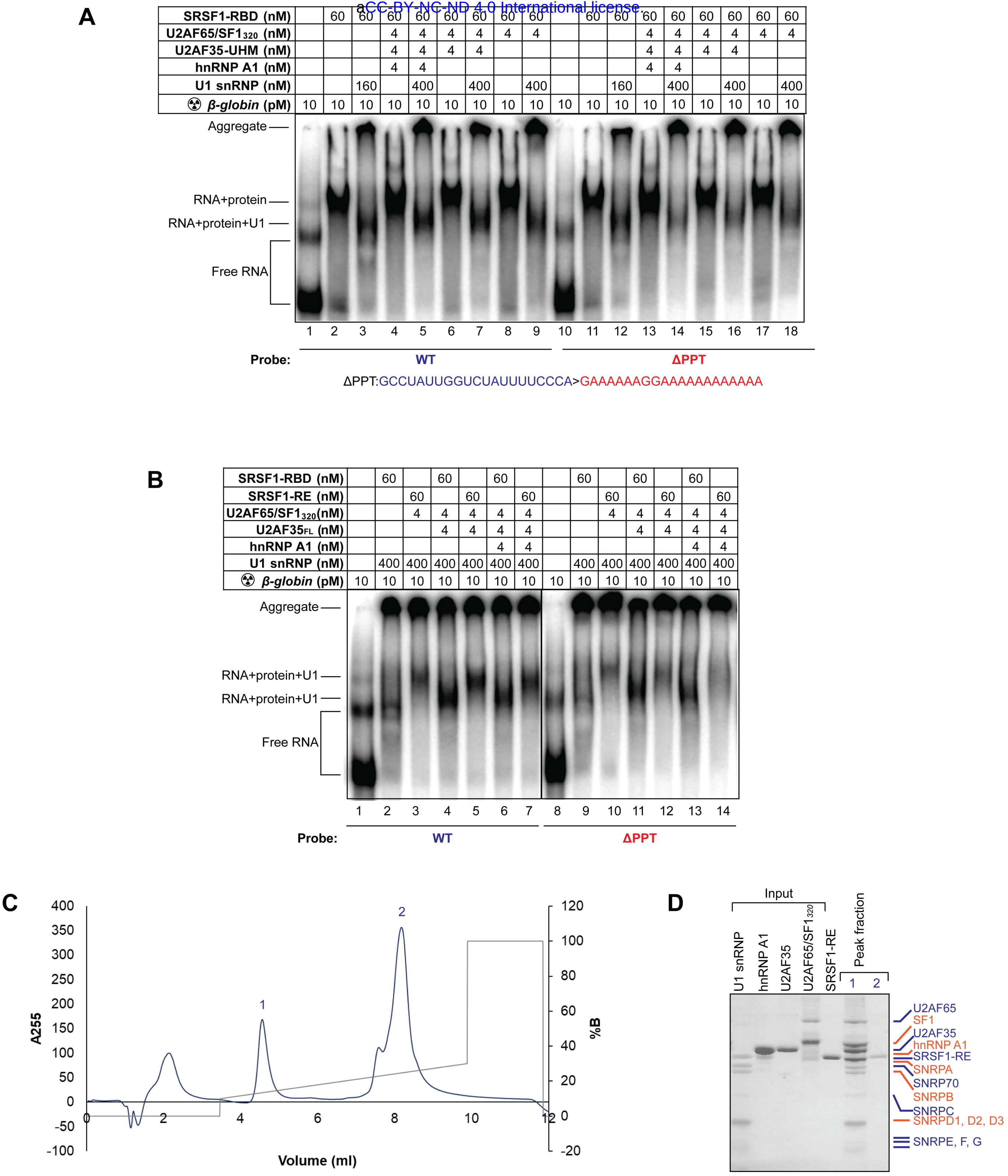
SRSF1-RBD (nM)		556	556						
SRSF1-RE (nM)				556	556				
SRSF1-RERA (nM)						556	556		
SRSF1-RS (nM)								556	556
U1 snRNP B$_{174}$ (μM)			1.5		1.5		1.5		1.5
β-globin (nM)	25	25	25	25	25	25	25	25	25
☢ **β-globin (pM)**	10	10	10	10	10	10	10	10	10

Figure S2



SRSF1-RBD (nM)		60	60	60	60	60	60	60	60	60	60	60
U1 snRNP (nM)			160	160	160	160	160	160	160	160	160	160
βg-5′SS (nM)			150									
DNA (μM)				0.8	1.6	2.4	3.2		0.8	1.6	2.4	3.2
RNase H (U)				1	1	1	1					
☢ *β-globin* (pM)	10	10	10	10	10	10	10	10	10	10	10	10

Figure S3



SRSF1-RBD (nM)		60	60	60	60	60	60	60	60		60	60	60	60	60	60	60	60
U2AF65/SF1$_{320}$ (nM)				4	4	4	4	4	4				4	4	4	4	4	4
U2AF35-UHM (nM)				4	4	4	4						4	4	4	4		
hnRNP A1 (nM)				4	4								4	4				
U1 snRNP (nM)			160		400		400		400			160		400		400		400
☢ *β-globin* (pM)	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

ΔPPT:GCCUAUUGGUCUAUUUUCCCA>GAAAAAAGGAAAAAAAAAAAA

SRSF1-RBD (nM)		60		60		60			60		60		60	
SRSF1-RE (nM)			60		60		60			60		60		60
U2AF65/SF1$_{320}$(nM)			4	4	4	4	4			4	4	4	4	4
U2AF35$_{FL}$ (nM)				4	4	4	4				4	4	4	4
hnRNP A1 (nM)						4	4						4	4
U1 snRNP (nM)		400	400	400	400	400	400		400	400	400	400	400	400
☢ *β-globin* (pM)	10	10	10	10	10	10	10	10	10	10	10	10	10	10

Figure S4