Transcriptome-wide profiling of mammalian spliceosome and branchpoints with iCLIP

Michael Briese[1,2]*, Nejc Haberman[3,4]*, Christopher R Sibley[1,4,5]*, Anob Chakrabarti[3,9], Zhen Wang[1], Julian König[1,6], David Perera[7], Vihandha O. Wickramasinghe[7,8], Ashok R. Venkitaraman[7], Nicholas M Luscombe[3,9,10], Christopher W. Smith[11], Tomaž Curk[12], Jernej Ule[1,3,4§]

[1]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge, CB2 0QH, UK

[2]Institute for Clinical Neurobiology, University of Wuerzburg, Versbacherstr. 5, 97078 Wuerzburg, Germany

[3]The Francis Crick Institute, Midland Rd, London, NW1 1AT, UK

[4]Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London, WC1N 3BG, UK

[5]Division of Brain Sciences, Department of Medicine, Imperial College London, London, WC12 0NN, UK

[6]Institute of Molecular Biology (IMB) gGmbH, Ackermannweg 4, 55128 Mainz, Germany

[7]The Medical Research Council Cancer Unit, University of Cambridge, Hills Road, Cambridge, CB2 0XZ, UK

[8]RNA Biology and Cancer Laboratory, Peter MacCallum Cancer Centre, 305 Grattan Street, Melbourne, Australia, 3000

[9]Department of Genetics, Environment and Evolution, UCL Genetics Institute, Gower Street, London WC1E 6BT, UK

[10]Okinawa Institute of Science & Technology Graduate University, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa 904-0495, Japan

[11]Department of Biochemistry, University of Cambridge, Downing Site, Tennis Court Road, Cambridge, CB2 1QW, UK

[12]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

Author Information:
Michael Briese, Nejc Haberman and Christopher R Sibley contributed equally to this work.

Corresponding author:
[§]Jernej Ule: jernej.ule@crick.ac.uk

Abstract

Studies of spliceosomal interactions are challenging due to their dynamic nature. Here we employed spliceosome iCLIP, which immunoprecipitates SmB along with snRNPs and auxiliary RNA binding proteins (RBPs), to simultaneously map human spliceosome engagement with snRNAs and pre-mRNAs. We identify nine sites on pre-mRNAs that overlap with position-dependent binding profiles of 15 RBPs. We reveal over 50,000 branchpoints (BPs), indicating that most human introns use a primary BP for spliceosome assembly, whereas alternative BPs are amplified when analysing intron lariats. Notably, we find that the binding patterns of many RBPs, especially the components of SF3 complex, are affected by RNA structure and position of BPs. Moreover, the stability of RNA structures around BPs distinguishes exons regulated by RBPs involved in early exon definition from those regulated by the SF3B components and PRPF8. These insights exemplify spliceosome iCLIP as a broadly applicable method for transcriptomic studies of BPs and splicing mechanisms.

53 **Introduction**

54 The vast majority of mammalian transcripts undergo splicing, a process through which
55 introns are excised and respective exons adjoined. Splicing is integral to gene expression,
56 whilst alternative splicing broadens the diversity of the transcriptome. It is a multi-step
57 process in which multiple snRNPs and associated splicing factors bind at specific
58 positions around intron boundaries in order to assemble an active spliceosome through
59 a series of remodeling steps. The splicing reactions are coordinated by dynamic pairings
60 between different snRNAs, between snRNAs and pre-mRNA, and by protein-RNA
61 contacts. Spliceosome assembly begins with ATP-independent binding of U1 snRNP at the
62 5'ss, as well as many auxiliary RBPs binding to pre-mRNA as part of complex E. This
63 includes the U2 small nuclear RNA auxiliary factors 1 and 2 (U2AF1 and U2AF2, also
64 known as U2AF35 and U2AF65) that bind at the 3'ss (Wahl et al., 2009). ATP-dependent
65 remodeling then leads to the formation of complex A. This enables U2 snRNP to contact
66 the BP, where it is stabilized through interactions with SF3a, SF3b and U2AF2. Next,
67 U4/U6 and U5 snRNPs are recruited to form complex B. The action of many RNA helicases
68 and pre-mRNA processing factor 8 (PRPF8) then facilitate rearrangements of snRNP
69 interactions and establishment of the catalytically competent B^{act} and C complexes. These
70 catalyze the two trans-esterification reactions that lead to lariat formation, intron
71 removal and exon ligation (Scheres and Nagai, 2017).

72 While techniques such as NET-seq and ribosome profiling monitor transcription or
73 translation through RNA protection by a single macromolecular machinery (Churchman
74 and Weissman, 2011; Ingolia et al., 2009), studies of splicing reaction are particularly
75 challenging due the multi-component and dynamic assembly of spliceosome on the pre-
76 mRNA substrate (Scheres and Nagai, 2017). Therefore, they have been studied mainly *in*
77 *vitro* with model substrates, or indirectly *in vivo* through monitoring the outcome of
78 splicing with techniques such as RNA-seq. More recently, high-throughput techniques to
79 study RNA interactions of the spliceosomal machinery have recently been developed
80 (Burke et al., 2018; Chen et al., 2018; Wickramasinghe et al., 2015). For studies in
81 mammalian cells, we have adapted the individual nucleotide resolution UV crosslinking
82 and immunoprecipitation (iCLIP) to developed spliceosome iCLIP, which defines the
83 positions of spliceosomal crosslinks on pre-mRNAs at nucleotide resolution
84 (Wickramasinghe et al., 2015). In yeast, "spliceosome profiling" has been developed
85 through affinity purification of tagged U2·U5·U6·NTC complex from *Schizosaccharomyces*
86 *pombe* to monitor its interactions with a RNA footprinting-based strategy (Burke et al.,
87 2018; Chen et al., 2018). It is currently unclear if this method can be applied to
88 mammalian cells, which might be more sensitive to the introduction of affinity tags into
89 splicing factors. Notably, none of these studies has examined the full complexity of the
90 interactions of diverse RBPs on pre-mRNAs from the earliest to the latest stages of
91 spliceosomal assembly.

92 A second challenge in understanding splicing mechanism is the need to assign the position
93 of branchpoints (BPs). The sequence consensus of mammalian BPs is less defined when
94 compared to yeast, and therefore experimental methods are important to assess
95 computational predictions. High-throughput methods to identify BPs have so far relied on

96 lariat-spanning RNA-seq reads that cross from the 5' portion of the intron, over the BP,
97 and finally finish in the 3' portion of the intron upstream of the BP (Mercer et al., 2015;
98 Pineda and Bradley, 2018; Taggart et al., 2017; Vogel et al., 1997). However, both the RNA-
99 seq methods and computational predictions tend to identify multiple potential BPs in
100 most introns, and it therefore remains unclear which BP is of primary functional
101 importance (Corvelo et al., 2010; Kol et al., 2005; Pagii and Bejerano, 2018; Pineda and
102 Bradley, 2018). Moreover, since lariat-spanning RNA-seq reads are very rare, they require
103 substantial sequencing depth. In yeast, spliceosome profiling was successful in assigning
104 the positions of BPs in yeast by monitoring the position of cDNAs truncating at BPs (Chen
105 et al., 2018), thus indicating that a similar approach could be applied also to mammalian
106 cells.

107 Here, we use spliceosome iCLIP to simultaneously monitor interactions of spliceosomal
108 RBPs across all the stages of the splicing cycle in human cell lines and mouse brain. We
109 find that this method can simultaneously map the crosslink profiles of core and accessory
110 spliceosomal factors that are known to participate across the diverse stages of splicing
111 cycle. In spite of this high diversity of purified factors, the nucleotide precision of iCLIP
112 reveals 7 binding peaks corresponding to distinct RBPs that differ in their requirement
113 for ATP or for the factor PRPF8. Spliceosome iCLIP also purifies intron lariats to identify
114 the positions of BPs. This allowed us to identify BPs in ~65% of introns within expressed
115 human genes, which contain more canonical sequence and structure features when
116 compared to those identified by RNA-seq. We have used the identified BPs to examine
117 surrounding binding of spliceosomal RBPs, which demonstrate that a single BP tends to
118 be primarily responsible for the assembly of SF3 and associated spliceosomal complex,
119 even though alternative BPs are detected by lariat-derived reads. Moreover, we identify
120 complementary roles of U2AF and SF3 complexes in BP definition; BPs with weak scores
121 are coupled with stronger binding of the U2AF complex, whilst BPs with higher scores
122 that tend to be presented within stem-loops are coupled with stronger binding and
123 function of SF3 complex. Taken together, these findings demonstrate the value of
124 spliceosome iCLIP for transcriptomic studies of BP definition and spliceosomal
125 interactions with pre-mRNAs.

126 **Results**

127 **Spliceosome iCLIP identifies interactions between splicing factors, snRNAs and pre-**
128 **mRNAs**

129 SmB/B' proteins are part of the highly stable Sm core common to all spliceosomal snRNPs
130 except U6 (Kambach et al., 1999), making them suitable candidates for enriching snRNPs
131 via immunopurification. In order to adapt iCLIP for the study of a multi-component
132 machine like the spliceosome, we immunopurified the SmB/B' proteins using the 18F6 or
133 12F5 monoclonal antibodies (Carissimi et al., 2006) (see online Methods for details)
134 (Figure 1A, Figure S1A-B), using a range of conditions with differing stringency. First, we
135 employed the standard condition used for iCLIP experiments in cultured human cells,
136 containing 0.1% SDS and all other detergents that lead to a stringent, but non-denaturing
137 purification (termed 'medium', Table S1). We examined the crosslinked radioactive

138 protein-RNA complexes on the SDS-PAGE gel upon use of a high RNase concentration,
139 which is optimized to create ~10 nt long RNA fragments (Huppertz et al., 2014). Two
140 strong radioactive bands were present, the first of which (at ~25 kDa) corresponds to the
141 molecular weight of the SmB-RNA complex (Figure 1B). In addition, a band at ~20 kDa
142 may correspond to another Sm protein, or another RBP that directly interacts with the
143 Sm ring. Next, we examined the radioactive protein-RNA complexes with the use of a
144 lower RNase concentration, which is optimized to create an average of 200 nt long RNA
145 fragments (with a full range of 50-500 nt) (Huppertz et al., 2014). Low RNase treatment
146 should allow snRNAs to remain generally intact, and therefore the snRNAs could serve as
147 a scaffold to purify a multi-protein complex (Figure 1A). In agreement, low RNase
148 treatment led to appearance of a diffuse signal between 30-200 kDa, indicating that
149 protein-RNA complexes of diverse sizes were present (Figure 1B). The same result was
150 obtained when the SmB/B' proteins were immunopurified using three different
151 monoclonal antibodies (Figure S1C).

152 Next, to examine if the signal between 30-200 kDa corresponds to the Sm proteins, or
153 other RBPs that are associated with snRNPs, we used a denaturing condition (termed
154 'stringent') that employed urea to specifically purify SmB via a Flag tag (Table S1). In
155 studies of other RBPs, this condition minimised co-purification of additional proteins
156 apart from that being purified (Huppertz et al., 2014) (Figure 1A). Indeed, we observed a
157 single band corresponding to the molecular weight of SmB-RNA complex, and the signal
158 remained largely unchanged under the low RNase condition (Figure S1D). This indicates
159 that purification of a multi-protein snRNP complex requires low RNase and purification
160 conditions that can preserve protein-protein interactions. Our standard iCLIP protocol
161 employs a high concentration of detergents in the lysis buffer, followed by washing buffer
162 with 1M NaCl (medium, Table S1), which is not compatible with many protein-protein
163 interactions, except the stable complexes such as snRNPs (Figure 1A). Therefore, we
164 established a modified purification with decreased concentration of detergents in the
165 lysis buffer, and only 0.1M NaCl in the washing buffer (termed 'mild', Table S1). Under this
166 mild purification condition, the diffuse signal at 30-200 kDa strongly increased upon low
167 RNase treatment, indicating that the mild purification allows most efficient purification
168 of multi-protein complexes, such as snRNPs and their associated proteins (Figure 1A,
169 Figure S1E).

170 To produce cDNA libraries with spliceosome iCLIP, we immunoprecipitated SmB under
171 conditions of three different levels of stringency from UV-crosslinked cell lysates, and
172 isolated a broad size distribution of protein-RNA complexes in order to recover the
173 greatest possible diversity of spliceosomal protein-RNA interactions (Figure 1B, S1C-E).
174 Lysates were treated with low RNase condition that leads to cDNAs of optimal size for
175 comprehensive crosslink determination (Haberman et al., 2017), and three biological
176 replicate samples of cDNA libraries were prepared (Table S2 and S3) (Huppertz et al.,
177 2014). As in previous iCLIP studies (König et al., 2010), the nucleotide preceding each
178 cDNA was considered as the crosslink site and cDNAs at each crosslink site were summed
179 as cDNA counts. When stringent conditions were used, >75% of spliceosomal crosslinking
180 mapped to snRNAs (Figure 1C), which agrees with its specific purification of Flag-tagged
181 SmB. However, upon mild and medium conditions of spliceosome iCLIP, the proportion

182 of snRNA crosslinking was reduced to approximately 10% and replaced by crosslinking
183 to introns and exons, indicating that these conditions allow mapping of snRNP-associated
184 proteins on pre-mRNAs (Figure 1A and C).

185 **Spliceosome iCLIP identifies seven crosslinking peaks on pre-mRNAs**

186 Assembly of the spliceosome on pre-mRNA is guided by three main landmarks: the 5'ss,
187 3'ss and BP. Therefore, we evaluated if spliceosomal crosslinks locate at specific positions
188 relative to these landmarks. The position of BPs was determined by using published
189 computationally predicted BPs (Pagii and Bejerano, 2018). Notably, when we plotted
190 summarized spliceosomal crosslinking at these landmarks of all introns, 7 peaks of
191 crosslinking were observed (Figure 2A, Figure S2A). In addition, two peaks precisely
192 align with the start of the intron and putative BPs. We term these positions A and B, and
193 these are discussed in detail in due course (Figure 2A, Figure S2A). The center of each
194 peak was seen at 15 nt upstream of the 5'ss (peak 1), the nucleotide at the start of the
195 intron (position A), 10 nt downstream of the 5'ss (peak 2), 31 nt downstream of the 5'ss
196 (peak 3), 26 nt upstream of the BP (peak 4), 20 nt upstream of the BP (peak 5), the
197 nucleotide of the BP (position B), 11 nt upstream of the 3'ss (peak 6) and 3 nt upstream
198 of the 3'ss (peak 7). The medium and mild conditions led to the same positions of the
199 peaks, with the signal of the mild condition generally being stronger, especially at the 3'ss
200 (Figure S2A). This is consistent with the stronger signal of co-purified complexes on the
201 SDS-PAGE gel under the mild condition (Figure S1D).

202 We used the mild condition to investigate how PRPF8 knockdown (KD) affects
203 spliceosomal interactions in Cal51 cells (Figure S2B). PRPF8 is an integral U5 snRNP
204 component, and therefore part of of complexes B and C, where it contacts residues of U5
205 and U6 snRNAs, as well as pre-mRNA at both the splice sites and BP (Teigelkamp et al.,
206 1995). We have previously used spliceosome iCLIP to show that PRPF8 KD affects the
207 crosslinking pattern of spliceosomal factors at the 5'ss (Wickramasinghe et al., 2015).
208 Here we additionally examined the crosslinking around BPs to find that PRPF8 KD
209 reproducibly leads to diminished crosslinking at peaks 1 and 3-5 (Figure 2A). Moreover,
210 we also observed a major decrease of reads truncating at the positions A and B, consistent
211 with the expectation that production of lariats should diminish upon PRPF8 KD. In
212 contrast, crosslinking at peaks 2 and 6 is increased upon PRPF8 KD. This demonstrates
213 that spliceosome iCLIP monitors pre-mRNA interactions of factors that interact with pre-
214 mRNAs at different stages of the splicing reaction.

215 ***in vitro* spliceosome iCLIP defines the ATP-dependence of crosslinking peaks**

216 Having identified the characteristic peaks of spliceosomal crosslinking *in vivo*, we then
217 tested if these peaks could also be observed under the controlled conditions of an *in vitro*
218 splicing reaction. To this end, we added an exogenous pre-mRNA splicing substrate to
219 HeLa nuclear extract in the presence or absence of ATP. The RNA substrate was produced
220 by *in vitro* transcription of a minigene construct containing a short intron and flanking
221 exons from the human C6orf10 gene. Gel electrophoresis analysis of splicing products
222 confirmed that ATP was required for the formation of intron lariats and other splicing

223 products (Figure S2C). We performed spliceosome iCLIP on the splicing reactions and
224 confirmed the presence of the spliced product with exon-exon junction reads, which were
225 specific to the library produced from extracts with ATP (+ATP) (Figure S2D-E, Table S4).
226 This indicates that spliceosome iCLIP is suitable to monitor progression of spliceosomal
227 assembly on pre-mRNAs during the splicing reaction.

228 We then visualized the crosslinking on the substrate RNA, and marked the positions of
229 peaks that corresponded best to those found on endogenous transcripts (Figure 2B). A
230 similar extent of crosslinking was seen at peaks 1, 2, 6 and 7 in the presence or absence
231 of ATP, indicating that these peaks correspond to ATP-independent contacts of early
232 spliceosomal factors. These crosslinks could originate from RBPs and snRNPs that are
233 capable of binding pre-mRNA in a pre-splicing H complex or the ATP-independent E
234 complex (Jurica et al., 2002) including TIA1/TIAL1 and U2AF2 proteins, which tend to
235 bind at positions corresponding to peaks 2 and 6, respectively (Bennett et al., 1992; Forch
236 et al., 2000). Both proteins bind U-rich sites in an ATP-independent manner, in agreement
237 with the presence of U-tracts at peaks 2 and 6. The peaks 4 and 5 located in the region 19-
238 40 nt upstream of the BP, and positions A and B, were most dependent on the presence of
239 ATP.

240 **Lariat-specific reads are efficiently obtained with spliceosome iCLIP**

241 Following crosslinking, the peptide that remains bound to the RNA after digestion of the
242 RBP can lead to termination of reverse transcription, and production of so called
243 'truncated cDNAs' (Lee and Ule, 2018). The predominance of truncated cDNAs in iCLIP
244 libraries has been validated by multiple means (Haberman et al., 2017; Sugimoto et al.,
245 2012), and therefore our analysis of iCLIP data refers to the nucleotide preceding the
246 iCLIP read on the reference genome as the 'crosslink site', and same applies to derived
247 methods, such as eCLIP. However, in the case of spliceosome iCLIP, we expect that the
248 presence of intron lariats that interact with splicing factors could lead to additional types
249 of truncated cDNAs. If an RBP is crosslinked in the vicinity of the BP, it could remain bound
250 to the three-way-junction that is formed upon fragmention by RNase, where the 5' end of
251 the intron is linked via a 2'-5' phosphodiester bond to the BP (Figure 2C). Such three-way-
252 junction RNAs present two available 3' ends for ligation to adapters, and the reads derived
253 from them can truncate at the BP (i.e. position B) or at the start of the intron (i.e. position
254 A), rather than representing the RBP crosslink site. In agreement with this hypothesis, we
255 find that cDNA truncations at positions A and B are dramatically decreased in conditions
256 where intron lariats are not expected to be present: *in vivo* upon PRPF8 KD (Figure 2A),
257 or *in vitro* in the absence of ATP (Figure 2B). Taken together, we find that *in vivo* depletion
258 of PRPF8 and *in vitro* depletion of ATP have similar effect on crosslinking of splicing
259 factors to pre-mRNAs, and also lead to loss of intron lariats. In particular, peaks upstream
260 of BPs are diminished upon both conditions, indicating that they represent interactions
261 of factors that are involved in the later stages of spliceosomal assembly.
262
263 **iCLIP identifies >50,000 human branchpoints (BPs) with canonical features**

264 Our analysis indicates that cDNAs truncating at BPs might result from the presence of
265 three-way junctions that cause termination of the reverse transcriptase at the BP (Figure
266 2C). We found that the medium purification condition was optimal to identify cDNAs that
267 truncate at peak B (Figure S2A), and we therefore performed spliceosome iCLIP under
268 medium purification conditions from UV crosslinked cell lysates (see Methods). Next, we
269 examined the nucleotide composition around the starts of cDNAs that end at the last
270 nucleotide of introns, which are most likely derived from intron lariats (see Figure 2C).
271 Interestingly, the starts of these cDNAs overlap with YUNAY motif, which is the known
272 consensus sequence of BPs (Sharp and Burge, 1997) (Figure 3A). Further, these cDNAs
273 have higher enrichment of mismatches of adenosines at their first nucleotide (Figure 3B),
274 which is consistent with mismatch, insertion and deletion errors during reverse
275 transcription across the three-way junction of the branch point (Mercer et al., 2015; Vogel
276 et al., 1997). In contrast, the starts of all remaining spliceosome iCLIP cDNAs overlap with
277 uridine-rich motif (Figure 3C), in agreement with the increased propensity of protein-
278 RNA crosslinking at uridine-rich motifs (Sugimoto et al., 2012). We conclude that cDNAs
279 overlapping with intron ends are mainly derived from intron lariats and truncate at BPs,
280 while the remaining cDNAs mainly represent RBP crosslink sites.

281 We used the spliceosome iCLIP cDNAs that overlap with intron ends to identify putative
282 BPs. We only used those cDNAs that start at adenines, which identified sites in 35,056
283 introns that may act as BPs. The more distal BPs would not be identified by this initial
284 approach due to our 41 read-length limit, and therefore we proceeded to a second step in
285 introns where the initial approach didn't identify any BPs. We analysed all cDNAs, and
286 overlapped their truncation sites with BPs computationally predicted in 2010 (Corvelo et
287 al., 2010). We selected the position with the highest number of truncated cDNAs, which
288 identified candidate BPs in another 15,756 introns. Thus, we collectively identified
289 candidate BPs in 50,812 introns of the most highly expressed genes (FPKM>10 as
290 determined by RNA-seq). Since these genes in total contain 78,894 annotated introns, we
291 were able to identify putative BPs in 64% of introns in expressed genes.

292 Next, we compared iCLIP BPs with studies that used lariat-spanning reads from RNA-seq
293 libraries to detect BPs with three different strategies; 59,359 BPs identified by
294 exoribonuclease digestion and targeted RNA-sequencing (Mercer et al., 2015), 36,078 BPs
295 identified by lariat-spanning reads refined by U2snRNP/pre-mRNA base-pairing models
296 (Taggart et al., 2017), or 130,294 BPs identified by analysis of lariat-spanning reads in
297 17,164 RNA sequencing data sets (Pineda and Bradley, 2018). In each case, we focused
298 on introns where a BP was defined both by the RNA-seq study and iCLIP, and we used
299 only a single BP per intron for each RNA-seq study (the one with most lariat-spanning
300 reads), to find a nucleotide resolution overlap with all studies (Figures S3A-B, 3D). 9348
301 (55% of Mercer), 6853 (45% of Taggart) and 11520 (54% of Pineda) BPs from each study
302 precisely overlapped with the BPs identified by spliceosome iCLIP (Figures S3C-D, 3E).

303 As a 1st step to compare the features of the non-overlapping BPs, where the positions of
304 BPs identified by RNA-seq ('RNA-seq BPs) doesn't overlap with those identified by
305 spliceosome iCLIP (iCLIP BPs), we examined the sequence consensus present at each class
306 of BPs. The overlapping BPs, as well the iCLIP BPs, were strongly enriched in the canonical

307 YUNAY motif (Figures S3E-F, S3H-I, 3F-G), while the RNA-seq BPs had poor enrichment
308 of the YUNAY motif, indicating that most are non-canonical (Figure S3G, S3J, 3H). As a 2nd
309 step, we examined the local RNA structure around each category of BPs (Figure 3I).
310 Notably, both the overlapping and the iCLIP-specific BPs had decreased intramolecular
311 pairing probability at the position of the BP. In contrast, the RNA-seq-specific BPs had the
312 opposite feature, with increased pairing probability in the region of the putative BP
313 (Figure 3I). This demonstrates that the RNA-seq-specific BPs contain mainly non-
314 canonical motifs that are in a structural conformation that is poorly accessible for pairing
315 with U2 snRNP. In contrast, spliceosome iCLIP identifies over 50,000 BPs that contain
316 canonical sequence and structural features.

317 As a 3rd step to compare the features of RNA-seq or iCLIP-specific BPs, we examined the
318 crosslinking of spliceosomal RBPs around the non-overlapping BPs. First, we analysed the
319 crosslink peaks identified by spliceosome iCLIP around the non-overlapping BPs (Figure
320 3J). The BPs identified by iCLIP were flanked by all the peaks at expected positions. This
321 not only included peak at position A, but also included peaks 4 and 5, even though the
322 cDNAs truncated in these peaks were not used to identify the BPs. In contrast, all the
323 peaks were shifted for BPs identified by RNA-seq – when these BPs located upstream of
324 iCLIP BPs, all peaks are shifted upstream of expected positions, and vice versa. We
325 consolidated this finding by using independent CLIP data that hasn't contributed to BP
326 identification: eCLIP and iCLIP data for PRPF8, SF3 and U2AF factors, which are known to
327 crosslink close to BPs (Figure 3K-L). Overlap of SF3 and U2AF crosslinking at the expected
328 positions of peaks 4-7 was good for the iCLIP-specific BPs (Figure 3K), but poor for RNA-
329 seq-specific BPs (Figure 3L). Moreover, the two groups of RNA-seq-specific BPs had their
330 peaks shifted by about 10 nt relative to each other. Thus, the RBPs binding profiles defined
331 by eCLIP or iCLIP in the Cal51, HeLa, K562 or HEPG2 cells do not align to the RNA-seq-
332 specific BPs, while the iCLIP-specific BPs align well to the binding profiles of these RBPs.
333 In spite of the lack of alignment between the RNA-seq-specific BPs and SF3/U2AF
334 complexes, we noticed that these BPs still partly aligned to the PRPF8 peak (Fig. 3M).

335 **Insights into introns with alternative branch points**

336 To gain further insight into the mechanisms that might contribute to the use of alternative
337 BPs, we next compared the iCLIP BPs with the BPs that have been bioinformatically
338 predicted. For this purpose, we used the most recent study that used a sequence-based
339 deep learning predictor, LaBranchoR, which was reported to predict a correct
340 branchpoint for over 90% of 3'ss (Pagii and Bejerano, 2018). Notably, the overlap
341 between iCLIP and computational BPs was higher than with RNA-seq BPs (Figure 4A). In
342 total, 31167 iCLIP BPs overlapped with the computationally top-scoring BP, while 11858
343 were over 5 nt apart from each other (Figure 4B). Only in 2393 of these non-overlapping
344 BPs, iCLIP-specific BPs overlapped with a lower ranking computational BP, indicating that
345 most of the iCLIP-specific BPs were not easy to predict computationally. Both the
346 overlapping and non-overlapping BPs have a very strong enrichment of the consensus
347 YUNAY motif (Figure 4C-E). Interestingly, the BPs that were up to 5nt apart tend to occur
348 within A-rich sequences (Figure S4A-B). Moreover, all groups of BPs had strikingly
349 decreased intramolecular pairing probability at the position of the BP, and high pairing

350 probability on both sides around the BP, indicative of a propensity for BP to be presented
351 on a loop within a stem-loop structure (Figure 4F-G). We noticed that the non-overlapping
352 BPs have an extended region of decreased pairing probability around the BPs, indicating
353 that they don't form such a clear stem loop. This indicates that a less defined RNA
354 structure at BPs may be a factor contributing to the use of alternative BPs. Interestingly,
355 the conservation of nucleotides is high immediately next to all types of BPs, while the
356 region surrounding the iCLIP-specific BPs is more highly conserved (Figure 4H).

357 Next, we examined the crosslinking of spliceosomal RBPs around the non-overlapping
358 BPs. Analysis of spliceosome iCLIP shows that the computation-specific BPs align better
359 to the spliceosomal peaks 4 and 5 than the iCLIP-specific BPs (Figure S4C). Moreover,
360 eCLIP and iCLIP data for SF3 and U2AF factors align better to the computational-specific
361 BPs (Figure 4J) than iCLIP-specific BPs (Figure 4I). In contrast, the PRPF8 eCLIP data align
362 with nucleotide precision to both sets of the non-overlapping BPs. As discussed earlier,
363 PRPF8 eCLIP cDNA starts most likely reflect the presence of reads derived from intron
364 lariats, which truncate at BPs (position B in Figure 2C). Thus, the computationally
365 predicted BPs appear to coordinate the assembly of U2 snRNP and the associated SF3
366 and U2AF complexes, while both classes of BPs contribute to lariat formation. This
367 indicates that most introns primarily use a single BP for spliceosome assembly, which can
368 be computationally predicted, while the lariat-based methods amplify the detection of
369 alternative BPs.

370 **The position of BPs defines the binding profiles of many spliceosomal RBPs**

371 Transcriptome-wide RNA maps represent valuable tools to study the position-dependent
372 binding profiles of many splicing factors (Ule et al., 2006; Witten and Ule, 2011). This
373 showed that function of RBPs is often determined by the position of their binding relative
374 to splice sites, but less is known about their orientation relative to BPs. We therefore
375 sought to identify the RBPs binding at peaks defined by the spliceosome iCLIP that are
376 aligned to BP (peaks 4, 5, B), and compare them to those binding at peaks aligned to 3'ss
377 (peaks 6 and 7). For this purpose, we examined published iCLIP data produced in our lab
378 for the 17 previously studied RBPs (Attig et al., 2018), and eCLIP data from K562 and
379 HepG2 cells for 112 RBPs provided by the ENCODE consortium (Van Nostrand et al.,
380 2017). For this analysis, we examined the introns containing the 31167 BPs that were
381 identified both computationally and by iCLIP, which are likely the most reliable. We
382 ranked the RBPs by the proportion of crosslink events mapping to each peak, and
383 normalized them by the combined region encompassing 200 nt of intronic sequence
384 upstream of the 3'ss and 50 nt of exonic sequence downstream of the 3'ss. Strikingly, a
385 different set of RBPs was enriched at each peak (Figure 5, Table S5).

386 According to previous studies, the region encompassing peaks 4 and 5 should represent
387 ATP-dependent binding of factors associated with the U2 snRNP complex, especially the
388 U2 snRNP splicing factor 3 (SF3a/SF3b) proteins, which were previously shown to
389 crosslink upstream of the BP (Gozani et al., 1996). Indeed, SF3B4, SF3A3 and SF3B1 are
390 among the factors with strongest enrichment of crosslinking in these peaks (Figure 5).
391 Surprisingly, we observed strong enrichment of additional RBPs in peaks 4 or 5, including

392 SMNDC1, GPKOW, BUD13, XRN2 and EFTUD2. Conversely, crosslinking of U2AF1 and
393 U2AF2 factors strongly dominates peaks 6 and 7, in agreement with their expected
394 binding patterns (Kielkopf et al., 2004; Zarnack et al., 2013). We validated that the U2AF
395 complex is co-purified in spliceosome iCLIP with western blot, especially in the absence
396 of ATP, which enriches for the complex E that is known to contain U2AF (Figure S2F).

397 To examine how the position of BPs on endogenous introns determines interactions of
398 spliceosomal RBPs, we divided 3'ss into three categories based on the distance between
399 BP and the 3'ss: 17-23 nt, 24-39 nt and 40-65 nt. We then analysed the position of cDNA
400 starts of the eCLIP and iCLIP data for the 13 RBPs that are most enriched in peaks 4-7, or
401 at position B. For each category of BPs we drew two diagrams; one where crosslinking
402 was summarized to the BP, and the other where crosslinking aligned to 3'ss (Figures S5,
403 S6). This confirmed that crosslinking of U2AF1 and U2AF2 is restricted to the region
404 between the BPs and 3'ss, while the RBPs binding to peaks 4 and 5 are precisely aligned
405 to the position of the BP, regardless of the distance between BP and 3'ss.

406 The crosslinking pattern of SF3B4 most closely reflected spliceosomal iCLIP at peaks 4
407 and 5 (Figure S5). SF3B4 contains two RRM domains, which might crosslink within the
408 positions of peaks 4 and 5. Interestingly, many additional RBPs have a binding pattern
409 similar to SF3B4, but with lesser enrichment. These again included SMNDC1, SF3B1,
410 EFTUD2 and XRN2 (Figure S5, S6). While it is possible that these RBPs independently bind
411 to pre-mRNA in a pattern similar to SF3B4, it is equally plausible that they are part of a
412 SF3B4-containing U2 snRNP complex that remains stable under the conditions of eCLIP.
413 In that case, the similar crosslinking pattern would reflect the presence of contaminating
414 SF3B4 in the eCLIP data of these other RBPs. At present, it is not possible to distinguish
415 between these two options, because the purified protein-RNA complexes need to be
416 visualized after their separation on SDS-PAGE gel to determine whether the expected
417 RBPs have been purified (Lee and Ule, 2018).

418 Several other RBPs at least partly overlapped with the spliceosomal peaks that align to
419 BPs. The crosslinking profile of SF3A3 has a peak 15 nt upstream of the BP (Figure S6).
420 Two RBPs that have very similar binding profile are BUD13 and GPKOW; their crosslinks
421 distribute equally between peaks 4 and 5. This suggests that these two RBPs either co-
422 purify as a complex, or independently crosslink in a very similar way upstream of BPs.
423 Finally, several RBPs had cDNA starts enriched at intron start (position A) and the BP
424 (position B): PRPF8, RBM22 and SUPV3L1 (Figure 5, S5, S6). PRPF8 and RBM22 are
425 known to associate with intron lariat as part of the human catalytic step I spliceosome
426 (Rasche et al., 2012; Scheres and Nagai, 2017; Teigelkamp et al., 1995), in agreement with
427 our hypothesis that both positions result largely from cDNAs truncating at the three-way
428 junction of intron lariats (Figure 2C). In conclusion, the RBPs binding in the region of
429 peaks 4 or 5 bind at a fixed distance relative to the position of the BP (Figure S5, S6),
430 indicating that their binding is defined primarily by the assembly of the spliceosome on
431 BPs, which acts as a molecular ruler that positions each associated RBP on pre-mRNAs at
432 a specific distance from BPs.

433 **RNA structure around BPs affects the binding and function of splicing factors**

434 To understand how the strength of spliceosomal RBP binding to pre-mRNAs is
435 determined by the features of BPs, we next summarised the crosslinking of RBPs by
436 normalising it relative to the complexity of the sequenced cDNA libraries, in order to
437 study the relative intensities of their binding (Figure 6). This shows that when BPs are
438 located close to the exon (Figure 6A), binding of U2AF factors at the 3'ss is weaker and
439 narrower compared to the exons with more distally located BPs (Figure 6C). This is
440 especially clear for U2AF2 binding to peak 6, which spans the whole region between the
441 3'ss and the BP. This can be explained by the characteristics of polypyrimidine (polyY)
442 tracts, which are precisely defined by the BP position, and span the region between the
443 BP and the AG dinucleotide of the 3'ss (Figure 6).

444 In contrast to the decreased binding of U2AF2 at proximal BPs, the binding of SF3
445 components increases at these BPs (Figure 6A). We were not able to identify any enriched
446 sequence motifs in the region of the peaks 4 or 5 (data not shown), and therefore the
447 intensity of SF3 binding is unlikely to be determined by the sequence properties of the
448 RNA. Thus, while binding of U2AF reflects the properties of polyY tracts, the intensity of
449 SF3 binding more likely reflects the recognition of BPs by the larger spliceosomal
450 complex. Thus, it appears that the stronger BP recognition may compensate for weaker
451 polyY-tract recognition by the U2AF complex.

452 To assess how the features of BPs affect the relative binding of SF3 and U2AF complexes,
453 we focused our further analyses on all the BPs identified by spliceosome iCLIP that are
454 located at 23-28 nt upstream of the intron-exon junction, which are the most common
455 positions of BPs (20018 BPs, Table S6). In this way we minimized the effects of variations
456 in BP position in order to focus on other features of BPs. Since we observed an effect of
457 BP position on the the relative binding of SF3/U2AF complexes (Figure 6), we examined
458 if the strength of BPs might also affect SF3/U2AF binding ratio. For this purpose, we
459 separated BPs into 10 equally sized quintiles based on their computationally predicted
460 score (Pagii and Bejerano, 2018). This revealed a significant trend, with stronger binding
461 of U2AF complex next to low-scoring BPs, and stronger binding of SF3 complex next two
462 high-scoring BPs, with an over 4-fold change in the ratio between the two complexes
463 when comparing the extreme quintiles ($p<0.001$, Wilcoxon Rank Sum test, Figure 7A).
464 Since the BP scores were defined by a sequence-based model, we checked if this change
465 in ratio could reflect a correlation between BP scores and the sequence of the polyY tract,
466 but we found that BP scores are independent of the proportion of Cs and Us within the
467 polyY tract (Figure S7A).

468 The BP scores were determined with a deep-learning model that is based purely on RNA
469 sequence, and therefore RNA structure was not used as one of the predictors.
470 Nevertheless, we examined the pairing probability of the nucleotides around the BPs
471 using RNAfold (Lorenz et al., 2011). We used only the 25nt region around the BPs as the
472 input for analysis, in order to evaluate the propensity of this region to form a local stem-
473 loop. Surprisingly, the pairing probability in the nucleotides surrounding the BPs was
474 strongly correlated with the BP scores, such that BPs with the highest scores had the
475 highest probability to form a stem-loop, where the BP is presented within the loop (Figure
476 S7B). As a result, the SF3/U2AF binding ratio is also correlated to the RNA structure

477 around BPs (data not shown). By shifting the position of the 25nt analysis window, we
478 find that the stronger pairing propensity is partly a feature of the sequence upstream of
479 BPs (Figure S7C-D). By analyzing the 10 quintiles of BP score, we confirmed that the
480 propensity to form a local stem loop is significantly increased for the high-scoring BPs
481 ($p<0.001$, Wilcoxon Rank Sum test, Figure 7B).

482 Given the link between RNA structure and RBP binding profiles, we wished to understand
483 whether the RNA structure around BPs also affects whether the exons are sensitive to
484 depletion of specific spliceosomal RBPs. For this purpose we used exons that become
485 more skipped upon knockdown of well-known splicing factors, as identified by the
486 ENCODE project from RNA-seq data following knockdown in K562 and HepG2 cells (Van
487 Nostrand et al., 2017). We only examined RBPs that play positive roles in splicing, either
488 by promoting exon definition during early stages of splicing (SR and U2AF proteins) or by
489 participating in spliceosomal assembly (SF3 and PRPF8 proteins). Given that these RBPs
490 play positive roles, we only examined the features of exons that are more skipped upon
491 knockdown. Notably, exons dependent on SF3B components or PRPF8 tend to contain
492 BPs that are presented within a stem loop, whereas exons dependent on factors involved
493 in exon definition (SRSF1 and U2AF component) tend to contain BPs with lower pairing
494 probability of their flanking nucleotides (Figure 7C). This effect is diminished if the
495 analysis window is shifted (Figure S7E-F), indicating that the local stem loop formation is
496 the primary factor. The groups of exons significantly differ in the pairing probability of
497 the nucleotides flanking the BPs ($p<0.01$, Wilcoxon Rank Sum test, Figure 7D), but not in
498 BP scores or Y-tract coverage (Figure S7G-H). In conclusion, we find that the strength of
499 the RNA structure around BPs affect the relative binding of SF3 and U2AF components,
500 and distinguishes exons regulated by RBPs involved in exon definition from those
501 regulated by the factors involved in spliceosome assembly on BPs, such as SF3B
502 components and PRPF8.

503

504 **Discussion**

505 We developed the spliceosome iCLIP method by using Sm core proteins as a bait to purify
506 endogenous snRNPs and their spliceosomal RBPs, as well as accessory splicing factors.
507 This enabled mapping their binding profiles along pre-mRNAs, which defined seven
508 primary binding peaks of spliceosomal factors around splice sites and BPs. Moreover, the
509 presence of lariat-derived reads in spliceosome iCLIP revealed >50,000 BPs that showed
510 good agreement with the latest computational predications. Notably, we find that the
511 computationally predicted BPs are primarily responsible for coordinating the assembly
512 of U2 snRNP and the associated SF3 and U2AF complexes, even though the lariat-based
513 methods often detect alternative BPs. By means of overlap with published iCLIP and eCLIP
514 data, we identified 15 RBPs that overlap with the spliceosomal peaks. The binding of most
515 of these RBPs is dependent on the position of BPs, and the corresponding spliceosomal
516 peaks are sensitive to ATP under *in vitro* conditions, and to PRPF8 knockdown *in vivo*.
517 Finally, we find that the strength of RNA structure around BPs affects the relative binding
518 of SF3 and U2AF complexes, and distinguishes exons regulated by RBPs involved in exon

519 definition (such as U2AF) from those involved in later stages of splicing reaction,
520 including the SF3B components and PRPF8.

521 We found that spliceosome iCLIP can also be used to identify BPs on a genome-wide scale
522 since it enriches for cDNAs derived from intron lariats that truncate at BPs. This identified
523 >50,000 BPs (64% of all expressed introns). For comparison, a recent RNA footprinting-
524 based strategy identified BPs in 26% (1420/5533) of all yeast introns using reads
525 spanning the BP three-way junction (Chen et al., 2018). This was improved to 86% when
526 BP characteristics were computationally inferred from both the lariat spanning reads and
527 all reads ending at ends of introns, although the precise number of unique BPs supported
528 by sequence reads remains unclear. Interestingly, BPs identified by iCLIP have a stronger
529 consensus sequence compared to the BPs previously experimentally identified in
530 mammalian cells through analysis of lariat reads in RNA-seq. The reason for this
531 difference may lie in the fact that lariats formed at non-canonical BPs are less efficiently
532 debranched, and thus they are likely to be more stable (Hartmuth and Barta, 1988). As a
533 result, detection of non-canonical BPs may be amplified by methods that are based on
534 analysis of lariats in RNA-seq due to the differential stability of lariats. Similarly, delayed
535 debranching of the non-canonical BPs could delay the kinetics of release of the late
536 spliceosomal complexes from intron lariats, thus explaining why PRPF8 eCLIP and
537 spliceosome iCLIP aligned better to the position of the non-canonical BPs than the SF3
538 crosslinking peaks 4 and 5, which correspond to the pre-lariat assembly of spliceosome
539 on BPs.

540 Spliceosome iCLIP compares favorably to other experimental methods, most likely
541 because it detects lariats that are in complex with spliceosome at time of crosslinking,
542 which should be less sensitive to variations in debranching kinetics and lariat stability.
543 We suggest that a combined of the lariat-derived reads in spliceosome iCLIP, the reads
544 that detect crosslinking in peaks 4 and 5, and computational modelling of BPs (Pagii and
545 Bejerano, 2018), could allow the most reliable identification and monitoring of the use of
546 functionally important BPs across development, tissues and species.

547 Interestingly, we observe that for the great majority of introns, the same BPs is identified
548 both computationally and by iCLIP. This class of BPs tend to be surrounded by high
549 pairing probability, indicating that they are often within a stem loop structure, with a
550 short loop that is enough just to present the consensus BP sequence. In contrast, the non-
551 overlapping BPs are more often surrounded by regions of extended low pairing
552 probability, indicating that the use of alternative BPs may result from their increased
553 accessibility for pairing with U2 snRNA. This would agree with the findings in yeast,
554 where conditions that change the structure of RNA (e.g. heat shock) allow new BPs to
555 become available to the spliceosome (Meyer et al., 2011). It will be important to test
556 whether RNA structure contributes to mammalian BP selection, and whether the choice
557 of alternative BPs that are identified by iCLIP or computational predictions can be
558 regulated under conditions that change RNA structure.

559 We found two peaks of spliceosomal crosslinking upstream of BPs. These perfectly
560 overlap with the crosslinking of SF3B4 and SF3B1, and partly with SF3A3, which are

561 required for the ATP-dependent step of spliceosome assembly on the BPs (Brosi et al.,
562 1993). This is in agreement with the ATP-dependence of peak 4 *in vitro* and its disruption
563 by PRPF8 KD. The binding positions of SF3B4 (peak at 26 nt upstream of BPs) and SF3A3
564 (peak at 15 nt upstream of BPs) is consistent with the structure of the human activated
565 spliceosome, where SF3A3 (also referred to as SF3a60) binds to pre-mRNA at a position
566 closer to the BP compared to SF3B4 (also referred to as SF3b49) (Zhang et al., 2018). The
567 SF3A3 peak does not overlap well with spliceosome iCLIP, indicating that it is not co-
568 purified with U2 snRNP as efficiently as SF3B4 under the conditions of spliceosome iCLIP.
569 Interestingly, while we observe binding peaks in the region 26-19 nt upstream of BPs in
570 humans, the late spliceosomal components in yeast had their peak centred at ~48-49 nt
571 upstream of BPs (Chen et al., 2018). In yeast, these contacts were predicted to correspond
572 to Cwf11, an SF1 RNA helicase (Chen et al., 2018). However, SF1 is expected to bind at a
573 distance of ~33-40 nucleotides from the BP based on structural studies (De et al., 2015),
574 and the identity of the yeast peak at 48-49 nt thus requires further experimental
575 validation. Irrespective, the contacts located upstream of the BPs are free of a clear motif
576 in both yeast and human. This implies that their binding position is defined by the
577 assembly of spliceosome on BPs, rather than by their individual RNA binding specificities.

578 In addition to the SF3 components, we found several other RBPs crosslinked at peaks 4/5
579 upstream of BPs. Among these are SMNDC1, GPKOW, BUD13 and EFTUD2, which have
580 been shown to regulate spliceosomal remodeling at the ATP-dependent stages of the
581 splicing cycle. SMNDC1, also known as SPF30, has originally been identified as a
582 component of the spliceosome by proteomic analyses (Neubauer et al., 1998) and has
583 similarity to the Survival Motor Neuron (SMN) protein (Talbot et al., 1998). Like SMN,
584 SMNDC1 contains a Tudor domain that binds to symmetrically dimethylated arginine
585 residues in the C-terminal tails of Sm proteins (Buhler et al., 1999). As part of the 17S U2
586 snRNP, SMNDC1 is essential for splicing and is involved in the transition of the pre-
587 spliceosomal complex A to the mature spliceosomal complex B (Meister et al., 2001).
588 GPKOW (Spp2 in yeast) has been identified as an interactor of the ATP-dependent RNA
589 helicase DHX16 (Prp2 in yeast) (Hegele et al., 2012; Roy et al., 1995; Silverman et al.,
590 2004). Spp2 and Prp2 are essential for the first splicing reaction and mediate the
591 transition of the pre-catalytic B^{act} to the catalytic B* complex (Kim and Lin, 1996;
592 Warkocki et al., 2015). Spp2/Prp2-mediated remodeling of the spliceosome leads to
593 destabilization of SF3a and SF3b proteins, which is thought to prepare the BP for the first
594 transesterification reaction (Lardelli et al., 2010; Warkocki et al., 2009). BUD13 has been
595 originally identified as a component of the trimeric RES (pre-mRNA REtention and
596 Splicing) complex in yeast, together with the proteins Snu17 and Pml1 (Dziembowski et
597 al., 2004). In the absence of the RES complex, spliceosomal B complexes are prematurely
598 disassembled by Prp2 to suggest that RES is required for an ordered remodeling of B^{act}
599 into B* complexes (Bao et al., 2017). Given the extremely similar crosslinking pattern of
600 GPKOW and BUD13 upstream of the BP, it will be interesting to explore if their functions
601 in the transition from B^{act} into B* complexes are related. Finally, EFTUD2 is a component
602 of spliceosomal C complex that interacts with PRPF8 (Achsel et al., 1998). Taken together,
603 the fact that that crosslinking pattern of spliceosome iCLIP overlaps with these diverse
604 RBPs indicates that it can simultaneously monitor pre-mRNA interactions of RBPs that
605 act at different stages of the splicing cycle.

606 Two RBPs, XRN2 and SUPV3L1, have not yet been reported to interact with spliceosome,
607 but their crosslinking profiles indicate their likely interactions. Xrn2 is a 5′→3′
608 exonuclease that associates with, and co-transcriptionally degrades, aberrantly
609 processed pre-mRNAs when processing is inhibited by Spliceostatin A (Davidson et al.,
610 2012). The crosslinking profile of XRN2 is most similar to SF3B4 and SF3B1, and,
611 interestingly, the SF3B complex is the target of Spliceostatin A (Kaida et al., 2007). It could
612 thus be hypothesized that XRN2 might detect aberrant splicing progression by contacting
613 the pre-mRNA at the same position that would normally be occupied by SF3B, thus
614 degrading pre-mRNAs that lack appropriate SF3B assembly. In the case of SUPV3L1, its
615 strong eCLIP signal at BPs indicates that it interacts with intron lariats. SUPV3L1 has not
616 yet been reported to have any functions in splicing, and has been studied solely as an ATP-
617 dependent RNA helicase within the mitochondrial exoribonuclease complex (Razew et al.,
618 2018). However, the protein is also found in cell nuclei, where it appears to regulate cell
619 cycle (Szewczyk et al., 2017), and thus a function in splicing remains plausible.

620 We show that BP position, and the computationally predicted strength of BPs, which is
621 highly correlated with RNA structure, affect the relative binding of U2AF and SF3
622 complexes upstream and downstream of BPs, respectively. Moreover, the strength of RNA
623 structure around BPs distinguishes exons regulated by PRPF8, SF3B or U2AF
624 components. This indicates complementary roles of these factors in spliceosome
625 assembly to ensure high fidelity of BP and 3'ss recognition. Interestingly, U2AF1, SF3B1
626 and PRPF8 are all targets for mutations in myeloid neoplasia (Kurtovic-Kozaric et al.,
627 2015; Yoshida et al., 2011). Even though these factors primarily bind at different regions
628 of introns (U2AF1 at 3'ss, SF3B1 upstream of BPs and PRPF8 at 5'ss), our study reveals
629 that the RNA secondary structure may contribute to the increased sensitivity of specific
630 introns to deficient function of these factors. Looking beyond, spliceosome iCLIP could be
631 used in the future to monitor transcriptome-wide impact of disease-causing mutations on
632 spliceosome assembly in human cells.

633 High-throughput analysis of RNA interactions of multi-protein complexes is crucial in
634 order to understand the dynamic assembly of such complexes on target RNAs. To date
635 this has been achieved for the human exon junction complex (EJC) with an RNA
636 footprinting method termed RNA:protein immunoprecipitation in tandem (RIPiT)-seq
637 (Singh et al., 2012), and the late-stage intron lariat spliceosome complex in yeast using
638 triple tagging of the U2.U5.U6.NTC complex (Burke et al., 2018; Chen et al., 2018). Both
639 approaches avoid UV-crosslinking in order to capture RBPs with poor UV crosslinkability
640 (Singh et al., 2014). However, those sites where RNase protection is inefficient may be lost
641 by these methods. Spliceosome iCLIP thus represents an alternative approach for the
642 study of multi-protein complexes with nucleotide resolution. Its capacity to monitor
643 concerted pre-mRNA binding of many types of spliceosomal proteins indicates that a
644 similar approach could be readily applied to other complexes composed of multiple RBPs
645 that bind RNA with distinct position-dependent binding patterns.

646
647

648

649 **Acknowledgements**

650 We thank Livio Pellizzoni for the 18F6 monoclonal antibody, Miriam Llorian for help with
651 the *in vitro* splicing reactions, Kathi Zarnack and Gregor Rot for help with the data
652 analyses, and Lisa Strittmatter, Ina Huppertz, Kiyoshi Nagai, Yoichiro Sugimoto, Jan Attig
653 and Martina Halleger for helpful discussions and comments on the manuscript. This work
654 was supported primarily by the European Research Council (206726-CLIP and 617837-
655 Translate) and the Slovenian Research Agency (P2-0209, Z7-3665, J7-5460). CRS is
656 supported by an Edmond Lily Safra fellowship. NML is a Winton Group Leader in
657 recognition of the Winton Charitable Foundation's support towards the establishment of
658 the Francis Crick Institute. NML is additionally funded by a Wellcome Trust Joint
659 Investigator Award (103760/Z/14/Z, the MRC eMedLab Medical Bioinformatics
660 Infrastructure Award (MR/L016311/1) and core funding from the Okinawa Institute of
661 Science & Technology Graduate University. DP and VOW were supported by Medical
662 Research Council programme grants MC_UU_12022/1 and MC_UU_12022/8 to ARV. The
663 Francis Crick Institute receives its core funding from Cancer Research UK
664 (FC001002), the UK Medical Research Council (FC001002), and the Wellcome Trust
665 (FC001002).

666 **Author contributions**

667 MB, CRS and JU conceived the project, designed the experiments and wrote the
668 manuscript, with assistance of all co-authors. MB, CRS and ZW performed experiments,
669 with assistance from JU, JK and CWS. NH performed most computational analyses, with
670 assistance from CRS, TC, AC and NML. VOW, DP and ARV provided crosslinked pellets
671 from wild-type and PRPF8-depleted Cal51 cells.

672 **Declaration of Interests**

673 The authors declare no competing interests.

674

675

676

677

678

679 **References:**

680 Achsel, T., Ahrens, K., Brahms, H., Teigelkamp, S., and Luhrmann, R. (1998). The human U5-
681 220kD protein (hPrp8) forms a stable RNA-free complex with several U5-specific proteins,
682 including an RNA unwindase, a homologue of ribosomal elongation factor EF-2, and a novel
683 WD-40 protein. Mol Cell Biol *18*, 6756-6766.

684 Attig, J., Agostini, F., Gooding, C., Chakrabarti, A., Singh, A., Haberman, N., Zagalak, J.,
685 Emmett, W., Smith, C.W., Luscombe, N.M.*, et al.* (2018). Heteromeric RNP assembly at LINEs
686 controls lineage-specific RNA processing. Cell *in press*.

687 Bao, P., Will, C.L., Urlaub, H., Boon, K.L., and Luhrmann, R. (2017). The RES complex is
688 required for efficient transformation of the precatalytic B spliceosome into an activated
689 B(act) complex. Genes Dev *31*, 2416-2429.

690 Bennett, M., Michaud, S., Kingston, J., and Reed, R. (1992). Protein components specifically
691 associated with prespliceosome and spliceosome complexes. Genes Dev *6*, 1986-2000.

692 Brosi, R., Hauri, H.P., and Kramer, A. (1993). Separation of splicing factor SF3 into two
693 components and purification of SF3a activity. J Biol Chem *268*, 17640-17646.

694 Buhler, D., Raker, V., Luhrmann, R., and Fischer, U. (1999). Essential role for the tudor
695 domain of SMN in spliceosomal U snRNP assembly: implications for spinal muscular atrophy.
696 Hum Mol Genet *8*, 2351-2357.

697 Burke, J.E., Longhurst, A.D., Merkurjev, D., Sales-Lee, J., Rao, B., Moresco, J.J., Yates, J.R.,
698 3rd, Li, J.J., and Madhani, H.D. (2018). Spliceosome Profiling Visualizes Operations of a
699 Dynamic RNP at Nucleotide Resolution. Cell *173*, 1014-1030 e1017.

700 Carissimi, C., Saieva, L., Gabanella, F., and Pellizzoni, L. (2006). Gemin8 is required for the
701 architecture and function of the survival motor neuron complex. J Biol Chem *281*, 37009-
702 37016.

703 Chen, W., Moore, J., Ozadam, H., Shulha, H.P., Rhind, N., Weng, Z., and Moore, M.J. (2018).
704 Transcriptome-wide Interrogation of the Functional Intronome by Spliceosome Profiling. Cell
705 *173*, 1031-1044 e1013.

706 Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes
707 transcription at nucleotide resolution. Nature *469*, 368-373.

708 Corvelo, A., Hallegger, M., Smith, C.W., and Eyras, E. (2010). Genome-wide association
709 between branch point properties and alternative splicing. PLoS computational biology *6*,
710 e1001016.

711 Davidson, L., Kerr, A., and West, S. (2012). Co-transcriptional degradation of aberrant pre-
712 mRNA by Xrn2. Embo J *31*, 2566-2578.

713 De, I., Bessonov, S., Hofele, R., dos Santos, K., Will, C.L., Urlaub, H., Luhrmann, R., and Pena,
714 V. (2015). The RNA helicase Aquarius exhibits structural adaptations mediating its
715 recruitment to spliceosomes. Nat Struct Mol Biol *22*, 138-144.

716 Dziembowski, A., Ventura, A.P., Rutz, B., Caspary, F., Faux, C., Halgand, F., Laprevote, O., and
717 Seraphin, B. (2004). Proteomic analysis identifies a new complex required for nuclear pre-
718 mRNA retention and splicing. Embo J *23*, 4847-4856.

719 Forch, P., Puig, O., Kedersha, N., Martinez, C., Granneman, S., Seraphin, B., Anderson, P., and
720 Valcarcel, J. (2000). The apoptosis-promoting factor TIA-1 is a regulator of alternative pre-
721 mRNA splicing. Mol Cell *6*, 1089-1098.

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

725 Haberman, N., Huppertz, I., Attig, J., Konig, J., Wang, Z., Hauer, C., Hentze, M.W., Kulozik,
726 A.E., Le Hir, H., Curk, T.*, et al.* (2017). Insights into the design and interpretation of iCLIP
727 experiments. Genome Biol *18*, 7.

728 Hartmuth, K., and Barta, A. (1988). Unusual branch point selection in processing of human
729 growth hormone pre-mRNA. Mol Cell Biol *8*, 2011-2020.

730 Hegele, A., Kamburov, A., Grossmann, A., Sourlis, C., Wowro, S., Weimann, M., Will, C.L.,
731 Pena, V., Luhrmann, R., and Stelzl, U. (2012). Dynamic protein-protein interaction wiring of
732 the human spliceosome. Mol Cell *45*, 567-580.

733 Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M.,
734 Konig, J., and Ule, J. (2014). iCLIP: protein-RNA interactions at nucleotide resolution.
735 Methods *65*, 274-287.

736 Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide
737 analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science
738 *324*, 218-223.

739 Jurica, M.S., Licklider, L.J., Gygi, S.R., Grigorieff, N., and Moore, M.J. (2002). Purification and
740 characterization of native spliceosomes suitable for three-dimensional structural analysis.
741 RNA *8*, 426-439.

742 Kaida, D., Motoyoshi, H., Tashiro, E., Nojima, T., Hagiwara, M., Ishigami, K., Watanabe, H.,
743 Kitahara, T., Yoshida, T., Nakajima, H.*, et al.* (2007). Spliceostatin A targets SF3b and inhibits
744 both splicing and nuclear retention of pre-mRNA. Nature chemical biology *3*, 576-583.

745 Kambach, C., Walke, S., Young, R., Avis, J.M., de la Fortelle, E., Raker, V.A., Luhrmann, R., Li,
746 J., and Nagai, K. (1999). Crystal structures of two Sm protein complexes and their
747 implications for the assembly of the spliceosomal snRNPs. Cell *96*, 375-387.

748 Kielkopf, C.L., Lucke, S., and Green, M.R. (2004). U2AF homology motifs: protein recognition
749 in the RRM world. Genes Dev *18*, 1513-1526.

750 Kim, S.H., and Lin, R.J. (1996). Spliceosome activation by PRP2 ATPase prior to the first
751 transesterification reaction of pre-mRNA splicing. Mol Cell Biol *16*, 6810-6819.

752 Kol, G., Lev-Maor, G., and Ast, G. (2005). Human-mouse comparative analysis reveals that
753 branch-site plasticity contributes to splicing regulation. Hum Mol Genet *14*, 1559-1568.

754 König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M.,
755 and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual
756 nucleotide resolution. Nat Struct Mol Biol *17*, 909-915.

757 Krainer, A.R., Maniatis, T., Ruskin, B., and Green, M.R. (1984). Normal and mutant human
758 beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. Cell *36*, 993-1005.

759 Kurtovic-Kozaric, A., Przychodzen, B., Singh, J., Konarska, M.M., Clemente, M.J., Otrock, Z.K.,
760 Nakashima, M., Hsi, E.D., Yoshida, K., Shiraishi, Y.*, et al.* (2015). PRPF8 defects cause
761 missplicing in myeloid malignancies. Leukemia *29*, 126-136.

762 Lardelli, R.M., Thompson, J.X., Yates, J.R., 3rd, and Stevens, S.W. (2010). Release of SF3 from
763 the intron branchpoint activates the first step of pre-mRNA splicing. Rna *16*, 516-528.

764 Lee, F.C.Y., and Ule, J. (2018). Advances in CLIP Technologies for Studies of Protein-RNA
765 Interactions. Mol Cell *69*, 354-369.

766 Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and
767 Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms for molecular biology : AMB *6*, 26.

768 Meister, G., Hannus, S., Plottner, O., Baars, T., Hartmann, E., Fakan, S., Laggerbauer, B., and
769 Fischer, U. (2001). SMNrp is an essential pre-mRNA splicing factor required for the formation
770 of the mature spliceosome. Embo J *20*, 2304-2314.

771 Melton, D.A., Krieg, P.A., Rebagliati, M.R., Maniatis, T., Zinn, K., and Green, M.R. (1984).
772 Efficient in vitro synthesis of biologically active RNA and RNA hybridization probes from
773 plasmids containing a bacteriophage SP6 promoter. Nucleic Acids Res *12*, 7035-7056.

774 Mercer, T.R., Clark, M.B., Andersen, S.B., Brunck, M.E., Haerty, W., Crawford, J., Taft, R.J.,
775 Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2015). Genome-wide discovery of human
776 splicing branchpoints. Genome Res *25*, 290-303.

777 Meyer, M., Plass, M., Perez-Valle, J., Eyras, E., and Vilardell, J. (2011). Deciphering 3'ss
778 selection in the yeast genome reveals an RNA thermosensor that mediates alternative
779 splicing. Mol Cell *43*, 1033-1039.

780 Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond,
781 A., and Mann, M. (1998). Mass spectrometry and EST-database searching allows
782 characterization of the multi-protein spliceosome complex. Nat Genet *20*, 46-50.

783 Pagii, J.M., and Bejerano, G. (2018). A sequence-based, deep learning model accurately
784 predicts RNA splicing branchpoints. bioRxiv.

785 Pineda, J.M.B., and Bradley, R.K. (2018). Most human introns are recognized via multiple and
786 tissue-specific branchpoints. Genes Dev *32*, 577-591.

787 Rasche, N., Dybkov, O., Schmitzova, J., Akyildiz, B., Fabrizio, P., and Luhrmann, R. (2012).
788 Cwc2 and its human homologue RBM22 promote an active conformation of the spliceosome
789 catalytic centre. EMBO J *31*, 1591-1604.

790 Razew, M., Warkocki, Z., Taube, M., Kolondra, A., Czarnocki-Cieciura, M., Nowak, E.,
791 Labedzka-Dmoch, K., Kawinska, A., Piatkowski, J., Golik, P*., et al.* (2018). Structural analysis
792 of mtEXO mitochondrial RNA degradosome reveals tight coupling of nuclease and helicase
793 components. Nature communications *9*, 97.

794 Roy, J., Kim, K., Maddock, J.R., Anthony, J.G., and Woolford, J.L., Jr. (1995). The final stages of
795 spliceosome maturation require Spp2p that can interact with the DEAH box protein Prp2p
796 and promote step 1 of splicing. Rna *1*, 375-390.

797 Scheres, S.H., and Nagai, K. (2017). CryoEM structures of spliceosomal complexes reveal the
798 molecular mechanism of pre-mRNA splicing. Curr Opin Struct Biol *46*, 130-139.

799 Sharp, P.A., and Burge, C.B. (1997). Classification of introns: U2-type or U12-type. Cell *91*,
800 875-879.

801 Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014).
802 rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-
803 Seq data. Proc Natl Acad Sci U S A *111*, E5593-5601.

804 Silverman, E.J., Maeda, A., Wei, J., Smith, P., Beggs, J.D., and Lin, R.J. (2004). Interaction
805 between a G-patch protein and a spliceosomal DEXD/H-box ATPase that is critical for
806 splicing. Mol Cell Biol *24*, 10101-10110.

807 Singh, G., Kucukural, A., Cenik, C., Leszyk, J.D., Shaffer, S.A., Weng, Z., and Moore, M.J.
808 (2012). The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR
809 protein nexus. Cell *151*, 750-764.

810 Singh, G., Ricci, E.P., and Moore, M.J. (2014). RIPiT-Seq: a high-throughput approach for
811 footprinting RNA:protein complexes. Methods *65*, 320-332.

812 Sugimoto, Y., Konig, J., Hussain, S., Zupan, B., Curk, T., Frye, M., and Ule, J. (2012). Analysis of
813 CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions.
814 Genome biology *13*, R67.

815 Szewczyk, M., Fedoryszak-Kuska, N., Tkaczuk, K., Dobrucki, J., Waligorska, A., and Stepien,
816 P.P. (2017). Human SUV3 helicase regulates growth rate of the HeLa cells and can localize in
817 the nucleoli. Acta biochimica Polonica *64*, 177-181.

818 Taggart, A.J., Lin, C.L., Shrestha, B., Heintzelman, C., Kim, S., and Fairbrother, W.G. (2017).
819 Large-scale analysis of branchpoint usage across species and cell lines. Genome Res *27*, 639-
820 649.

821 Talbot, K., Miguel-Aliaga, I., Mohaghegh, P., Ponting, C.P., and Davies, K.E. (1998).
822 Characterization of a gene encoding survival motor neuron (SMN)-related protein, a
823 constituent of the spliceosome complex. Hum Mol Genet *7*, 2149-2156.

824 Teigelkamp, S., Newman, A.J., and Beggs, J.D. (1995). Extensive interactions of PRP8 protein
825 with the 5' and 3' splice sites during splicing suggest a role in stabilization of exon alignment
826 by U5 snRNA. EMBO J *14*, 2602-2612.

827 Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B.J.,
828 and Darnell, R.B. (2006). An RNA map predicting Nova-dependent splicing regulation. Nature
829 *444*, 580-586.

830 Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Blue, S.M., Dominguez, D., Cody,
831 N.A.L., Olson, S., Sundararaman, B.*, et al.* (2017). A Large-Scale Binding and Functional Map
832 of Human RNA Binding Proteins. bioRxiv.

833 Vogel, J., Hess, W.R., and Borner, T. (1997). Precise branch point mapping and quantification
834 of splicing intermediates. Nucleic Acids Res *25*, 2030-2031.

835 Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: design principles of a
836 dynamic RNP machine. Cell *136*, 701-718.

837 Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N.M., Rot, G., Zupan, B., Curk, T.,
838 and Ule, J. (2010). iCLIP predicts the dual splicing effects of TIA-RNA interactions. PLoS Biol *8*,
839 e1000530.

840 Warkocki, Z., Odenwalder, P., Schmitzova, J., Platzmann, F., Stark, H., Urlaub, H., Ficner, R.,
841 Fabrizio, P., and Luhrmann, R. (2009). Reconstitution of both steps of Saccharomyces
842 cerevisiae splicing with purified spliceosomal components. Nat Struct Mol Biol *16*, 1237-
843 1243.

844 Warkocki, Z., Schneider, C., Mozaffari-Jovin, S., Schmitzova, J., Hobartner, C., Fabrizio, P., and
845 Luhrmann, R. (2015). The G-patch protein Spp2 couples the spliceosome-stimulated ATPase
846 activity of the DEAH-box protein Prp2 to catalytic activation of the spliceosome. Genes Dev
847 *29*, 94-107.

848 Wickramasinghe, V.O., Gonzalez-Porta, M., Perera, D., Bartolozzi, A.R., Sibley, C.R.,
849 Hallegger, M., Ule, J., Marioni, J.C., and Venkitaraman, A.R. (2015). Regulation of constitutive
850 and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5'
851 splice site strength. Genome Biol *16*, 201.

852 Witten, J.T., and Ule, J. (2011). Understanding splicing regulation through RNA splicing maps.
853 Trends Genet.

854 Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-
855 Otsubo, A., Kon, A., Nagasaki, M.*, et al.* (2011). Frequent pathway mutations of splicing
856 machinery in myelodysplasia. Nature *478*, 64-69.

857 Zarnack, K., Konig, J., Tajnik, M., Martincorena, I., Eustermann, S., Stevant, I., Reyes, A.,
858 Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct Competition between hnRNP C and
859 U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. Cell *152*, 453-466.

860 Zhang, X., Yan, C., Zhan, X., Li, L., Lei, J., and Shi, Y. (2018). Structure of the human activated
861 spliceosome in three conformational states. Cell research *28*, 307-322.

862

863

Legends:

Figure 1. Spliceosome iCLIP identifies protein interactions with snRNAs and splicing substrates

(A) Schematic representation of the spliceosome iCLIP method performed under conditions of varying lysis stringency.

(B) Autoradiogram of crosslinked RNPs immunopurified with SmB/B' from HeLa cells following digestion with high (++) or low (+) amounts of RNase I in medium-stringency lysis buffer. The dotted line depicts the region excised from the nitrocellulose membrane for spliceosome iCLIP.

(C) Genomic distribution of spliceosome iCLIP cDNAs.

Figure 2. Analysis of splicesomal interactions with pre-mRNAs in vitro and in vivo.

(A) RNA map of summarised crosslinking at all exon-intron and intron-exon boundaries detected by spliceosome iCLIP from Cal51 cells to identify major binding peaks, and to monitor changes between the WT and PRPF8 KD. Individual lines of comparable colour represent replicate datasets.

(B) Normalized spliceosome iCLIP cDNA counts on the C6orf10 in vitro splicing substrate. Exons are marked by grey boxes, intron by a line, and the BP by a green dot. The positions of peak crosslinking are marked by numbers and letters corresponding to the peaks in Figure 2A.

(C) Schematic description of the three-way junctions of intron lariats. The three-way junction is produced after limited RNase I digestion of intron lariats, followed by ligation of the L3 adaptor to the two 3' ends of the three-way junction. This can lead to cDNAs that don't truncate at sites of protein-RNA crosslinking, but rather at the three-way junction of intron lariats. These cDNAs initiate from the end of the intron and truncate at the BP (position B), or initiate downstream of the 5' splice site and truncate at the first nucleotide of the intron (position A).

Figure 3. Comparison of BPs identified by spliceosome iCLIP or RNA-seq lariat reads.

(A) Weblogo around the nucleotide preceding all spliceosome iCLIP reads that align with ends of introns.

(B) Enrichment of mismatches at the first nucleotide of spliceosome iCLIP reads that overlap with ends of introns, compared to remaining iCLIP reads.

(C) Weblogo around the nucleotide preceding all remaining spliceosome iCLIP reads (i.e., those that do not align with ends of introns).

(D) The number of top BPs identified by published RNA-seq (Pineda and Bradley, 2018) (i.e., the BP with most lariat-spanning reads in each intron) at positions relative to BPs identified by spliceosome iCLIP (i.e., iCLIP BPs).

(E) A table providing the number of BPs identified by spliceosome with iCLIP (iCLIP BPs) in introns that also contain a BP assigned by lariat-spanning reads from RNA-seq (Pineda and Bradley, 2018). They are divided into three categories based on the distance between the iCLIP BP and the top RNA-seq BP.

905 (F) Weblogo of iCLIP BPs that overlap with RNA-seq BPs.

906 (G) Weblogo of iCLIP BPs that are >5 nt away from RNA-seq BP.

907 (H) Weblogo of RNA-seq BPs that are >5 nt away from iCLIP BP.

908 (I, J) The 80 nucleotide RNA region centered on the BP was used to calculate pairing
909 probability with RNAfold program with the default parameters (Lorenz et al., 2011), and
910 the average pairing probability of each nucleotide around BPs was calculated for each
911 category of BPs.

912 (K) RNA map of summarised crosslinking of spliceosome iCLIP from Cal51 cells around
913 each category of BPs.

914 (L) Crosslinking patterns of selected RBPs, as defined by cDNA starts of eCLIP in the
915 indicated cell lines, around iCLIP BPs that are located upstream (dotted line) or
916 downstream (filled line) of the RNA-seq BPs.

917 (M) Crosslinking patterns of selected RBPs, as defined by cDNA starts of eCLIP in the
918 indicated cell lines, around RNA-seq BPs that are located upstream (dotted line) or
919 downstream (filled line) of the iCLIP BPs.

920

921 **Figure 4. Comparison of computationally predicted BPs with those identified with**
922 **spliceosome iCLIP.**

923 (A) The number of top computationally predicted BPs (Pagii and Bejerano, 2018) (i.e., the
924 top scoring BP in each intron) at positions relative to BPs identified by spliceosome iCLIP
925 (i.e., iCLIP BPs).

926 (B) A table providing the number of BPs identified by spliceosome with iCLIP (iCLIP BPs)
927 in introns that also contain a computationally predicted BP (Pagii and Bejerano, 2018).
928 They are divided into three categories based on the distance between the iCLIP BP and
929 the top computational BP.

930 (C) Weblogo of iCLIP BPs that overlap with comp BP.

931 (D) Weblogo of iCLIP BPs that are >5 nt away from comp BP.

932 (E) Weblogo of comp BPs that are >5 nt away from iCLIP BP.

933 (F, G) The 80 nucleotide RNA region centered on the BP was used to calculate pairing
934 probability with RNAfold program with the default parameters (Lorenz et al., 2011), and
935 the average pairing probability of each nucleotide around BPs was calculated for each
936 category of BPs, as defined by the legend.

937 (H) Average mammalian conservation score of each nucleotide around BPs was calculated
938 for each category of BPs, as defined by the legend.

939 (I) Crosslinking patterns of selected RBPs, as defined by cDNA starts of eCLIP in the
940 indicated cell lines, around iCLIP BPs that are located upstream (dotted line) or
941 downstream (filled line) of the computational BPs.

942 (J) Crosslinking patterns of selected RBPs, as defined by cDNA starts of eCLIP in the
943 indicated cell lines, around computational BPs that are located upstream (dotted line) or
944 downstream (filled line) of the iCLIP BPs.

945

946 **Figure 5. Identification of RBPs overlapping with splicesomal peaks at BPs and 3'ss.**

947 To systematically identify RBPs with crosslinking peaks that overlap with one of the peaks
948 in spliceosomal iCLIP, we regionally normalized (relative to average crosslinking over -
949 200..50 nt region relative to 3'ss) the crosslinking of each RBP, as shown in Fig. S5 and S6.
950 We then calculated the average normalized crosslinking across the nucleotides within
951 each peak, and ranked the RBPs relative to this normalized crosslinking enrichments. We
952 analysed peaks 4-7, which are located close to 3'ss, and positions A and B, as marked on
953 the top of each plot. The plot of top-ranking RBPs in each peak is shown on the left, and
954 the full distribution of RBP enrichments is shown on the right.

955

956 **Figure 6. BP position defines the binding patterns of splicing factors at 3'ss.**

957 (A) Crosslinking patterns of selected RBPs, as defined by cDNA starts of eCLIP or iCLIP in
958 the indicated cell lines. All 3'ss that contain BPs within 17..23 nt upstream of the exon are
959 chosen, and crosslinking is plotted in the region -40..10 nt relative to 3'ss, and -40..10 nt
960 relative to BPs. Below the plot, a weblogo shows nucleotide composition in the same
961 region.

962 (B) Same as (A), but for all 3'ss that contain BPs within 24..39 nt upstream of the exon.

963 (C) Same as (A), but for all 3'ss that contain BPs within 40..65 nt upstream of the exon.

964

965 **Figure 7. RNA structure around BPs affects the binding and function of splicing**
966 **factors.**

967 (A) BPs were divided into 10 quantiles based on their sequence consensus score, as
968 determined in the past study (Pagii and Bejerano, 2018). The median score of each
969 quantile is shown on the x-axis. The 4410 BPs chosen for this analysis satisfied two
970 criteria: 1) They were located 23-28 nt away from intron-exon junction (the most
971 common locations of BPs), and 2) they contained a total of at least 30 crosslink events of
972 SF3B4−K562−eCLIP, SF3B4−HepG2−eCLIP, SF3A3−HepG2−eCLIP in the region 35-10 nt
973 upstream of BPs and U2AF2−HepG2−eCLIP, U2AF2−K562−eCLIP and
974 U2AF1−K562−eCLIP in the region 5-25 nt downstream of BPs (the peak binding region of
975 these RBPs). The y-axis shows the ratio in binding of SF3 relative to U2AF factors around
976 the BP, which was calculated by dividing the total cDNA count of SF3B4−K562−eCLIP,
977 SF3B4−HepG2−eCLIP, SF3A3−HepG2−eCLIP in the region 35-10 nt upstream of BPs with
978 the total cDNA count of U2AF2−HepG2−eCLIP, U2AF2−K562−eCLIP and
979 U2AF1−K562−eCLIP in the region 5-25 nt downstream of BPs. P-values for the indicated
980 comparisons were calculated by the pairwise Wilcoxon Rank Sum test.

981 (B) BPs were divided into 10 categories based on their sequence consensus score, as
982 determined in the past study (Pagii and Bejerano, 2018). The average pairing probability
983 of each nucleotide around BPs was then calculated for each category of BPs. The 80
984 nucleotide RNA region centered on the BP was used to calculate pairing probability with
985 RNAfold program using the default parameters (Lorenz et al., 2011). P-values for the
986 indicated comparisons were calculated by the pairwise Wilcoxon Rank Sum test.

987 (C) Exons that become more skipped upon knockdown of the indicated RBPs were
988 identified from data provided by the ENCODE project from RNA-seq data following
989 CRISPR knockdown in K562 and HepG2 cells (Van Nostrand et al., 2017). The
990 differentially spliced exons were detected with rMATS using the junction counts only, a
991 P-value threshold of 0.05, FDR threshold of 0.1 and IncLevelDifference > 0.2 (Shen et al.,
992 2014), and only exons with increased skipping upon knockdown were used. We used all
993 data where at least 45 skipped exons were detected with our threshold. The top
994 computationally predicted BPs were used for each set of exons, and the average pairing
995 probability was calculated for each nucleotide around BPs for each category of BPs. The
996 25 nucleotide RNA region around the BP was used to calculate pairing probability with
997 RNAfold program using the default parameters (Lorenz et al., 2011).

998 (D) The distribution of pairing probability for each group of exons that are skipped upon
999 knockdown of the indicated RBPs. The pairing probability was calculated as in C, and the
1000 average pairing probability at positions 12-7 nt upstream of BPs and 3-6 nt downstream
1001 of BPs was calculated.

1002

1003 **Supplementary legends**

1004 **Figure S1. Related to Figure 1: Quality control of spliceosome iCLIP with the anti-**
1005 **SmB/B' antibodies**

1006 (A) Western blot analysis of total HeLa cell extract with 18F6 antibody reveals a single
1007 band of 28 kDa.

1008 (B) Analysis of HeLa cells by immunostaining with 18F6 and epifluorescence
1009 microscopy shows expected localization of SmB/B' (a speckled nuclear pattern
1010 excluding nucleoli).

1011 (C) Autoradiograms of crosslinked RNPs after immunopurification with the anti-SmB/B'
1012 monoclonal antibodies 18F6 (as hybridoma supernatant), 12F5 (Carissimi et al., 2006)
1013 (sc-130670, Santa Cruz Biotechnology) or Y12 (ab3138, Abcam). HeLa cell pellet was
1014 lysed in medium lysis buffer and subjected to high (++, final dilution 1:10,000) or low (+,
1015 final dilution 1:100,000) concentrations of RNase I. Lysates were split evenly between
1016 beads for immunopurification. RNAs of immunopurified RNP complexes were
1017 radiolabelled at the 5' end followed by size-separation on denaturing gels and
1018 nitrocellulose transfer. The time below each panel indicates length of exposure during
1019 autoradiography.

1020 (D) UV-crosslinked HEK FLP-in cells with SmB-3XFLAG stably integrated were lysed
1021 under stringent conditions and subjected to partial RNase I digestion (Low - final
1022 dilution 1:100,000, High – dilution 1:5000). Spliceosomal RNPs were immunopurified
1023 with anti-FLAG M2 antibody, RNA was 5' end radiolabeled, and RNPs were subjected to
1024 denaturing gel electrophoresis and nitrocellulose transfer, an autoradiogram of which is
1025 shown. The interrupted line indicates the area on the nitrocellulose membrane cut out
1026 for purification of crosslinked RNP complexes.

1027 (E) UV-crosslinked mouse postnatal day 7 brains were lysed under medium or mild
1028 stringency conditions and subjected to partial RNase I digestion (final dilution
1029 1:100,000). Spliceosomal RNPs were immunopurified with anti-SmB/B' 18F6 antibody,

1030 RNA was 5' end radiolabeled, and RNPs were subjected to denaturing gel
1031 electrophoresis and nitrocellulose transfer, an autoradiogram of which is shown in the
1032 upper panel. The interrupted line indicates the area on the nitrocellulose membrane cut
1033 out for purification of crosslinked RNP complexes. For Western blotting, the remainder
1034 of the supernatant following cell lysis and centrifugation was mixed with 4× loading
1035 buffer (Invitrogen) and equal sample volumes were separated by SDS-PAGE and
1036 transferred onto nitrocellulose membrane, which was incubated with anti-α-tubulin
1037 antibody (1:4,000, clone B-5-1-2, cat. no. T5168, Sigma-Aldrich).

1038

1039 **Figure S2. Related to Figure 2: Analysis spliceosome iCLIP from cell extracts and in**
1040 **vitro splicing reactions.**

1041 (A) RNA map of summarized crosslinking for spliceosome iCLIP performed under
1042 medium or mild conditions from mouse brain is shown around the exon-intron, intron-
1043 exon junction and computationally top-scoring BP in each mouse intron (Corvelo et al.,
1044 2010).

1045 (B) Immunoblot analysis of Prpf8 knockdown efficiency in Cal51 cells.

1046 (C) RNAs transcribed in vitro from C6ORF10 minigene construct were incubated with
1047 HeLa nuclear extracts as part of in vitro splicing reactions in the presence or absence of
1048 ATP. Resulting splicing products and intermediates were resolved by denaturing gel
1049 electrophoresis and visualized by autoradiography.

1050 (D) In vitro splicing reactions were diluted in mild lysis buffer, subjected to low RNase I
1051 treatment (final dilution 1:200,000) and used for spliceosome iCLIP. Autoradiogram of
1052 crosslinked size-separated RNP complexes show the radiolabelled RNA that is
1053 crosslinked to RBPs. The interrupted line indicates the area cut out from the
1054 nitrocellulose membrane for extraction of crosslinked RNAs, which were used as a
1055 template for the iCLIP cDNA libraries.

1056 (E) U2AF65 and U2AF35 are co-purified with Sm proteins in the Spliceosome iCLIP.
1057 HeLa nuclear extracts were incubated for 1 hour with or without ATP and subjected to
1058 RNase I treatment (+, final dilution 1:5,000) where indicated followed by
1059 immunoprecipitation using SmB/B' antibody-conjugated protein-G beads. Indicated
1060 proteins were detected by immunoblotting.

1061 (F) Normalised spliceosome iCLIP cDNA counts on the C6orf10 in vitro splicing
1062 substrate. Exons are marked by grey boxes. As expected, junction reads are almost
1063 exclusively present only in the +ATP library.

1064

1065 **Figure S3. Related to Figure 3: Comparison of BPs determined by spliceosome**
1066 **iCLIP or RNA-seq lariat reads.**

1067 (A) The number of top BPs identified by published RNA-seq (Mercer et al., 2015) (i.e.,
1068 the BP with most lariat-spanning reads in each intron) at positions relative to BPs
1069 identified by spliceosome iCLIP (i.e., iCLIP BPs).

1070 (B) The number of top BPs identified by published RNA-seq (Taggart et al., 2017) (i.e.,
1071 the BP with most lariat-spanning reads in each intron) at positions relative to BPs
1072 identified by spliceosome iCLIP (i.e., iCLIP BPs).

1073 (C) A table providing the number of BPs identified by spliceosome with iCLIP (iCLIP BPs)
1074 in introns that also contain a BP assigned by lariat-spanning reads from RNA-seq . They
1075 are divided into three categories based on the distance between the iCLIP BP and the top
1076 RNA-seq BP.

1077 (D) A table providing the number of BPs identified by spliceosome with iCLIP (iCLIP BPs)
1078 in introns that also contain a BP assigned by lariat-spanning reads from RNA-seq (Taggart
1079 et al., 2017). They are divided into three categories based on the distance between the
1080 iCLIP BP and the top RNA-seq BP.

1081 (E) Weblogo of iCLIP BPs that overlap with RNA-seq BPs (Mercer et al., 2015).

1082 (F) Weblogo of iCLIP BPs that are >5 nt away from RNA-seq BP (Mercer et al., 2015).

1083 (G) Weblogo of RNA-seq BPs (Mercer et al., 2015) that are >5 nt away from iCLIP BP.

1084 (H) Weblogo of iCLIP BPs that overlap with RNA-seq BPs (Taggart et al., 2017).

1085 (I) Weblogo of iCLIP BPs that are >5 nt away from RNA-seq BP (Taggart et al., 2017).

1086 (J) Weblogo of RNA-seq BPs that are >5 nt away from iCLIP BP (Taggart et al., 2017).

1087

1088 **Figure S4. Related to Figure 4: Comparison of BPs determined by spliceosome**
1089 **iCLIP or computational lariat reads.**

1090 (A) Weblogo of iCLIP BPs that are >5 nt upstream or downstream of comp BP.

1091 (B) Weblogo of comp BPs that are >5 nt upstream or downstream of iCLIP BP.

1092 (C) RNA map of summarised crosslinking of spliceosome iCLIP from Cal51 cells around
1093 each category of BPs.

1094

1095 **Figure S5. Related to Figure 5: Crosslinking of many RBPs overlaps with a peak of**
1096 **spliceosomal crosslinking.**

1097 (A) Crosslinking patterns of selected RBPs, as defined by cDNA starts of eCLIP or iCLIP in
1098 the indicated cell lines. All 3'ss that contain BPs within 17..23 nt upstream of the exon are
1099 chosen, and crosslinking is plotted in the region -40..10 nt relative to 3'ss, and -40..10 nt
1100 relative to BPs. Crosslinking of each is regionally normalized (relative to its average
1101 crosslinking over -200..50 nt region relative to 3'ss).

1102 (B) Same as (A), but for all 3'ss that contain BPs within 24..39 nt upstream of the exon.

1103 (C) Same as (A), but for all 3'ss that contain BPs within 40..65 nt upstream of the exon.

1104

1105 **Figure S6. Related to Figure 5: Crosslinking of many RBPs overlaps with a peak of**
1106 **spliceosomal crosslinking.**

1107 (A) Crosslinking patterns of selected RBPs, as defined by cDNA starts of eCLIP or iCLIP in
1108 the indicated cell lines. All 3'ss that contain BPs within 17..23 nt upstream of the exon are
1109 chosen, and crosslinking is plotted in the region -40..10 nt relative to 3'ss, and -40..10 nt
1110 relative to BPs. Crosslinking of each RBP is regionally normalized (relative to its average
1111 crosslinking over -200..50 nt region relative to 3'ss).

1112 (B) Same as (A), but for all 3'ss that contain BPs within 24..39 nt upstream of the exon.

1113 (C) Same as (A), but for all 3'ss that contain BPs within 40..65 nt upstream of the exon.

1114

1115 **Figure S7. Related to Figure 7: RNA structure and sequence characterisation around**
1116 **BPs.**

1117 (A) BPs were divided into 10 quantiles as described in Figure 7A, with the median score
1118 of each quantile shown on the x-axis. The % of Ys (C or T) in the region 1-21 nt
1119 downstream of BP is shown on the y-axis.

1120 (B) BPs were divided into 10 quantiles as described in Figure 7A, and the 25 nucleotide
1121 RNA region centered on the BP was used to calculate pairing probability with RNAfold
1122 program with the default parameters (Lorenz et al., 2011), and the average pairing
1123 probability of each nucleotide around BPs is shown for each quantile.

1124 (C-D) Same as B, but the 25 nucleotide RNA region is shifted 10nt upstream or
1125 downstream of BPs.

1126 (E-F) Same as Figure 7C, but the 25 nucleotide RNA region is shifted 10nt upstream or
1127 downstream of BPs.

1128 (G) The distribution of computational BP score for each group of exons, which are skipped
1129 upon knockdown of the indicated RBPs.

1130 (H) The distribution of % of Ys (C or T) in the region 1-21 nt downstream of BP for each
1131 group of exons, which are skipped upon knockdown of the indicated RBPs.

1132

1133

1134

1135 **Methods:**

1136 **Data and statistics**

1137 The spliceosome iCLIP data have been deposited on EBI ArrayExpress under the
1138 accession number E-MTAB-6950. These and published data sets referenced throughout
1139 this study are listed for convenience in Table S7, including accession details. All statistical
1140 analyses were performed in the R software environment (version 3.1.3 and 3.3.2,
1141 https://www.r-project.org).

1142 **Preparation of Cal51 cells for iCLIP**

1143 Cal51 breast adenocarcinoma cells were prepared as described previously
1144 (Wickramasinghe et al., 2015). Briefly, cells were cultured in Dulbecco's Modified Eagle
1145 Medium (DMEM, Thermo Fisher) with 10% fetal calf serum (FCS, Thermo Fisher) and 1×
1146 penicillin-streptomycin (P/S, Thermo Fisher). For siRNA-mediated depletion of PRPF8,
1147 Cal51 cells were transfected with DharmaFECT1 (Dharmafect) with 25 nM siRNA against
1148 human PRPF8. Transfected cells were harvested 54 hrs later, exposed to UV-C light and
1149 used for iCLIP as described below. For collection of samples from different stages of the
1150 cell cycle, Cal51 cells were synchronized in G1/S by standard double thymidine block.
1151 Briefly, cells were treated with 1.5 mM thymidine for 8 hrs, washed and released for 8 hrs,
1152 then treated again with thymidine for a further 8 hrs. Cells were also collected 3 hrs (S-
1153 phase) and 7 hrs (G2) after release from the thymidine block.

1154 ***In vitro* splicing**

1155 For *in vitro* splicing reactions, we produced a C6orf10 minigene construct containing exon
1156 8 and 9 and 150 nt of the intron around both splice sites (Figure 2B). The minigene was
1157 linearized and transcribed *in vitro* using T7 polymerase with ^{32}P-UTP (Melton et al.,
1158 1984). The transcribed RNA was then subjected to *in vitro* splicing reactions using HeLa
1159 nuclear extract as described before (Krainer et al., 1984). Briefly, HeLa nuclear extract
1160 was depleted of endogenous ATP by pre-incubation and, for each reaction, 10 ng of RNA
1161 was incubated with 60% HeLa nuclear extract at 30°C with or without additional 0.5 mM
1162 ATP for 1 h in a 20 µl reaction. Afterwards, the reaction mixture was UV-crosslinked at
1163 100 mJ/cm^2 and stored at -80°C until further use. To visualise the splicing reaction
1164 products, proteinase K was added to the reaction mixture for 30 min at 37°C. The
1165 resulting RNA was phenol-extracted, precipitated and visualised on a 5% polyacrylamide-
1166 urea gel. To test the effect of ATP on U2AF protein association with the spliceosome, *in*
1167 *vitro* splicing reactions were set up without minigene substrate RNA in the presence or
1168 absence of 0.5 mM ATP. Reactions were incubated at 30°C for 1 h and then diluted to 1 ml
1169 using mild lysis buffer. As indicated, reactions were subjected to high RNase I (AM2295,
1170 100 U/µl, Thermo Fisher) treatment (+, final dilution 1:5,000) followed by
1171 immunoprecipitation using 12F5-conjugated protein-G beads. Indicated proteins were
1172 detected by Western blotting using anti-U2AF2 (sc-48804, Santa Cruz Biotechnology),
1173 anti-U2AF35 (sc-19961, Santa Cruz Biotechnology) or Y12 antibody (ab3138, Abcam)

1174 visualized using both horseradish peroxidase (HRP)-conjugated protein A (101123,
1175 Thermo Fisher) and protein G (101223, Thermo Fisher).

1176 **Spliceosome iCLIP protocol**

1177 The iCLIP method was done as previously described (König et al., 2010), with the
1178 following modifications. Crosslinked cells or tissue were dissociated in the lysis buffer
1179 according to the stringency conditions (stringent, medium, mild; Table S1) followed by
1180 sonication, limited RNase I digestion and centrifugation. For initial experiments, HEK293
1181 FlpIn cells expressing 3×Flag-SmB were used for high stringency conditions (Huppertz et
1182 al., 2014) (Figure S1D) and mouse brain tissue was used for testing mild and medium
1183 stringency conditions (Figure S1E). Subsequently, HeLa nuclear extract was used for *in*
1184 *vitro* splicing reactions (Figure S2C), and Cal51 cells used for medium and mild conditions.
1185 Identification of protein crosslink sites around splice sites, in particular the peak 4, is most
1186 efficient under the mild purification condition (Figure S2A). This condition was therefore
1187 used for spliceosome iCLIP from PRPF8 knockdown and control Cal51 cells (Figure 2A),
1188 and for *in vitro* splicing in HeLa nuclear extract (Figure 2B). For the identification of BPs,
1189 we found the medium condition is best suited, since it increases the frequency of cDNAs
1190 truncating at peak B (Figure S2A). Therefore, spliceosome iCLIP was performed under
1191 medium purification conditions from Cal51 cells synchronized in G1, S and G2 phase, and
1192 data from these libraries was pooled with the data from untreated cells produced under
1193 mild conditions (which also contains many cDNA starts overlapping with BPs, see Figure
1194 2A) in order to identify BPs.

1195 For SmB protein immunopurification under stringent conditions, M2 anti-Flag antibody
1196 (Sigma) was used against the 3×Flag-SmB protein that had been stably integrated into
1197 HEK-293 FlpIn cells. Briefly, immunopurification was performed as outlined previously
1198 (Huppertz et al. 2014). Here, a 6M Urea buffer was first used to lyse cell pellets, before
1199 being diluted down 1:9 with a Tween-20 containing IP buffer to allow for immuno-
1200 purification without denaturing of the M2 anti-Flag antibody. The iCLIP protocol was then
1201 followed as described previously. For SmB/B' immunopurification under medium and
1202 mild conditions anti-SmB/B' antibodies 12F5 and 18F6 were used, which are different
1203 clones from the same immunization, as described previously (Carissimi et al., 2006).
1204 These antibodies behave identically under immunopurification conditions (Figure S1C).
1205 For spliceosome iCLIP from mouse brain and *in vitro* splicing reactions, lysates were
1206 incubated with 50 µl monoclonal anti-SmB/B' antibody 18F6 [as hybridoma supernatant,
1207 generated as described previously (Carissimi et al., 2006)], and for experiments from
1208 Cal51 cells, 12F5 anti-SmB/B' antibody (Santa Cruz) was used. The antibody was pre-
1209 conjugated to 100 µl protein G Dynabeads (Thermo Fisher) and rotated at 4°C followed
1210 by washing. As described previously, following immunopurification, RNA 3' end
1211 dephosphorylation, ligation of the linker 5'-rAppAGATCGGAAGAGCGGTTCAG/ddC/-3' to
1212 the 3' end and 5' end radiolabeling protein-RNA complexes were size-separated by SDS-
1213 PAGE and transferred onto nitrocellulose membrane. The regions corresponding to 28-
1214 180 kDa were excised from the membrane in order to isolate the bound RNA by
1215 proteinase K treatment. RNAs were reverse-transcribed in all experiments using
1216 SuperScript III reverse transcriptase at U/µl (ThermoFisher, USA) and custom indexed

1217 primers (Table S2). Resulting cDNAs were subjected to electrophoresis on a 6% TBE-urea
1218 gel (Thermo Fisher) for size selection. Purified cDNAs were circularized, linearized and
1219 amplified for high-throughput sequencing. U2AF2 and TIA1 iCLIP data used in RNA maps
1220 were produced from HeLa cells in previous studies (Wang et al., 2010; Zarnack et al.,
1221 2013).

1222 **Mapping of Sm iCLIP reads**

1223 We used mm9/NCBI37 and hg19/GRCh37 genome versions sequence information and
1224 Ensembl 75 gene annotation. Experimental and random barcode sequences of iCLIP
1225 sequenced reads were removed prior to mapping (Table S2). We mapped the cDNAs to
1226 the genome with Bowtie 0.12.7 program using the parameters (-v 2 -m 1 -a --best --strata).
1227 The first 9 nt of the sequenced contains the experimental barcode to separate
1228 experimental replicates, and the UMIs, which allow to avoid artefacts caused by variable
1229 PCR amplification of different cDNAs (König et al., 2010). We used these UMIs to quantify
1230 the number of unique cDNAs that mapped to each position in the genome by collapsing
1231 cDNAs with the same UMI that mapped to the same starting position to a single cDNA. For
1232 analysis of crosslinking to snRNAs, we allowed sequences to map at up to 50 locations in
1233 the genome, but for all other analyses in the manuscript, we only allowed sequence
1234 mapping to a single location in the genome. The nucleotide preceding the iCLIP cDNAs
1235 mapped by Bowtie was used to define the crosslink sites.

1236 **Mapping of eCLIP reads**

1237 For eCLIP sequencing data for all RBPs, we used GENCODE (GRCh38.p7) genome
1238 assembly and the STAR alignment (version 2.4.2a) using the following parameters from
1239 ENCODE pipeline: STAR --runThreadN 8 --runMode alignReads --genomeDir GRCh38
1240 Gencode v25 --genomeLoad LoadAndKeep --readFilesIn read1, read2, --
1241 readFilesCommand zcat --outSAMunmapped Within –outFilterMultimapNmax 1 --
1242 outFilterMultimapScoreRange 1 --outSAMattributes All --outSAMtype BAM Unsorted –
1243 outFilterType BySJout --outFilterScoreMin 10 --alignEndsType EndToEnd --
1244 outFileNamePrefix outfile.

1245 For the PCR duplicates removal, we used a python script 'barcode collapse pe.py' available
1246 on GitHub (https://github.com/YeoLab/gscripts/releases/tag/1.0), which is part of the
1247 ENCODE eCLIP pipeline (https://www.encodeproject.org/pipelines/ENCPL357ADL/).

1248 **Visualization of crosslink positions in the form of RNA maps**

1249 RNA maps were produced by summarizing the cDNA counts at significant crosslink sites
1250 at all exon/intron and intron/exon boundaries and BPs on pre-mRNAs. The definition of
1251 intronic start and end positions was based on Ensembl version 75. Only introns longer
1252 than 300 nts were used to draw RNA maps, since they enabled best normalization of data
1253 by the intronic abundance with the following procedure:

1254 1. For each intron, calculate the total count of cDNAs that identify crosslink sites in the
1255 deep intronic region (from 50 nt downstream from exon-intron junction to 100 nt
1256 upstream of exon-intron junctions). If the count is >10, then proceed with the analysis of
1257 this intron. The average cDNA count/nt in the deep intronic regions is used as a
1258 normalisation constant.

1259 2. Divide the counts at each position in the intron and flanking exons by the normalisation
1260 constant of this intron.

1261 3. Sum up the normalised values from step 2 in the examined introns, and divide this value
1262 by the number of examined introns.

1263 For all analyses in Cal51 cells, we only assessed protein-coding genes with FPKM>10 in
1264 RNA-seq data. For spliceosome iCLIP with the C6orf10 *in vitro* splicing substrate,
1265 sequence reads were first mapped to the unspliced substrates and the remaining reads
1266 were mapped to the spliced substrate allowing no mismatches.

1267 **Identification of branchpoints (BPs)**

1268 To identify BPs, we pooled spliceosome iCLIP data produced under mild and medium
1269 purification conditions from Cal51 cells. The first step to identify BPs used the
1270 spliceosome iCLIP reads that ended precisely at the ends of introns (we considered only
1271 introns that end in AG dinucleotide) after removal of the 3' adapter. We noticed that these
1272 reads had an 3.5× increased frequency of mismatches on the A as the first nucleotide
1273 compared to remaining iCLIP reads (Figure 3B), indicating that these mismatches may
1274 have resulted from truncation at the three-way-junction formed at the BP (Figure 2C). We
1275 therefore trimmed the first nucleotide from the read if it contained a mismatch at the first
1276 position that corresponded to a genomic adenosine. We then used spliceosome iCLIP from
1277 synchronized and unsynchronized control Cal51 cells to identify all reads that ended
1278 precisely at the ends of introns and defined the position where these reads started and
1279 assessed the random nucleotides that are present at the beginning of each iCLIP read to
1280 count the number of unique cDNAs at each position. The nucleotide preceding the read
1281 start corresponds to the position where cDNAs truncated during the reverse
1282 transcription, and we selected the genomic A that had the highest number of truncated
1283 cDNAs as the candidate BP. If two positions with equal number of cDNAs were found, we
1284 selected the one closer to the 3' SS. This identified 35,056 positions in genes with
1285 FPKM>10. The sequence composition around these positions corresponded to the
1286 previously reported sequence around BPs (Figure 3A). We then proceeded to the second
1287 step of the analysis, where we considered all cDNAs (regardless of where they ended), but
1288 including trimming of the first nucleotide if there was a mismatch with the genomic A. We
1289 then overlapped cDNA truncation sites with computationally predicted BPs in the last 100
1290 nts of intron (Corvelo et al., 2010). If this analysis identified a position with a higher cDNA
1291 count than the initial analysis (or if the initial analysis didn't identify any BP in the same
1292 intron), then the newly identified position was assigned as the BP. For introns where no
1293 BP was identified by either first or second step in the analysis, we assessed
1294 computationally predicted BPs located further than 100 nts from the 3'ss, and if any of

1295 these overlapped with a truncating cDNA, we assigned the position closest to the 3'ss as
1296 the BP. Together, this identified 50,812 BPs in genes with FPKM>10. These BPs were used
1297 for all analyses in the manuscript, and their coordinates were used for BP positioning in
1298 RNA maps. We additionally identified 13,496 BPs in introns of lowly expressed genes, but
1299 these were not used for any further analyses.

1300 On RNA maps, the position of BPs is marked as peak B. This peak was more efficiently
1301 identified with spliceosome iCLIP on the endogenous pre-mRNAs when medium
1302 conditions were used, but was greatly diminished when mild conditions were used
1303 (Figure S2A). We hypothesize that the Dbr1 enzyme can debranch the lariats when the
1304 mild lysis conditions are used, and therefore cDNAs do not truncate at BPs under these
1305 conditions. In contrast, position B was detected on the exogenous substrate in spite of
1306 using mild conditions for spliceosome iCLIP (Figure 2B). We hypothesize that this was
1307 because the *in vitro* splicing extract lacks the Dbr1 enzyme, and therefore the lariats are
1308 not debranched during the iCLIP assay.

1309 Bedtools Intersect command using option –u was used to compare BP coordinates from
1310 spliceosome iCLIP to the BPs identified in previous studies. We restricted this comparison
1311 to introns where BPs were detected by all three datasets.

1312 **Comparison with the computationally predicted branch points**

1313 The top scoring BP positions predicted for each intron in hg19 were obtained for 3'ss
1314 (recommended) 8/25/2017) from (Pagii and Bejerano, 2018):

1315 http://bejerano.stanford.edu/labranchor/

1316 The iCLIP BP chosen for the comparison in each intron was the one containing the highest
1317 number of iCLIP reads truncating at the BP.

1318 **Analysis of pairing probability**

1319 Computational predictions of the secondary structure were performed by RNAfold
1320 function from Vienna Package (https://www.tbi.univie.ac.at/RNA/) with default
1321 parameters (Lorenz et al., 2011). The RNAfold results are provided in a customised
1322 format, where brackets are representing the double stranded region on the RNA and dots
1323 are used for unpaired nucleotides. We measured the density of pairing probability by
1324 summing the paired positions into a single vector.

1325 **Identification of RBPs overlapping with splicesomal peaks**

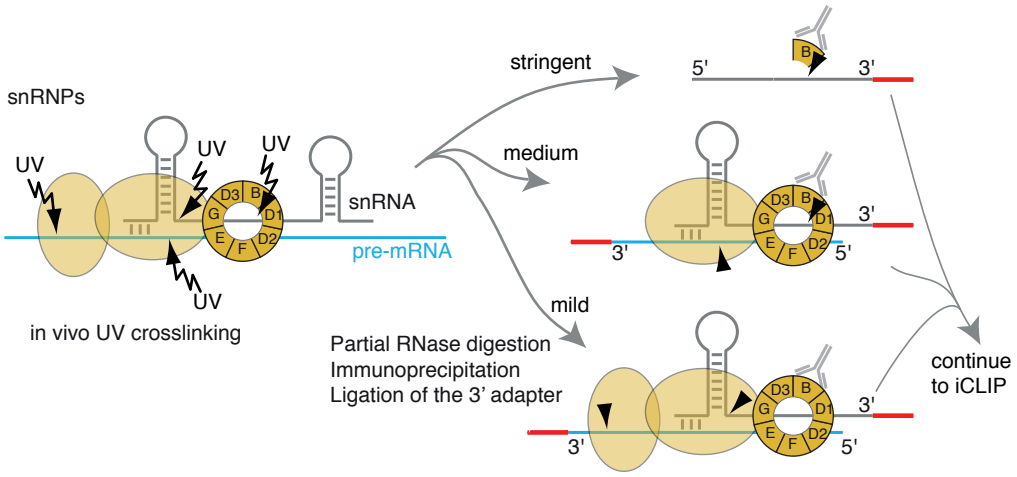
1326 For RBP enrichment in Figure 5, we used the eCLIP data from the ENCODE consortium
1327 (Van Nostrand et al., 2017), together with available iCLIP experiments from our lab (Attig
1328 et al., 2018), to see if any of the proteins are enriched in the region of splicesomal peaks.
1329 In total this included 140 eCLIP samples of 70 RBPs in the HepG2 cell line, and 178 eCLIP

1330 samples of 89 RBPs in the K562 cell, and iCLIP samples of 18 RBPs from different cell lines
1331 (Table S5). Next, we intersected cDNA-starts from each sample to the -200 to +50 nt
1332 region relative to the relative splice site and used it as control for each of the following
1333 peaks: Peak 4 (-23nt..-29nt relative to BP), Peak 5 (-21nt..-17nt relative to BP), Peak B (-
1334 1nt..1nt relative to BS) , Peak A (-1nt..1nt of 5'ss), Peak6 (-10nt..-11nt), Peak7 (-3nt..-2nt).
1335 The positions of these peaks were determined based on crosslink enrichments in
1336 spliceosome iCLIP.

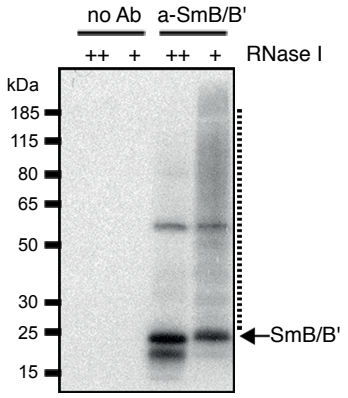
1337

Figure 1

A


snRNPs
UV
UV
UV
UV
snRNA
pre-mRNA
in vivo UV crosslinking
stringent
medium
mild
5'
3'
3'
5'
3'
5'
3'
5'
Partial RNase digestion
Immunoprecipitation
Ligation of the 3' adapter
continue
to iCLIP

B


no Ab
a-SmB/B'
++ + ++ +
RNase I
kDa
185
115
80
65
50
30
25
15
SmB/B'

C

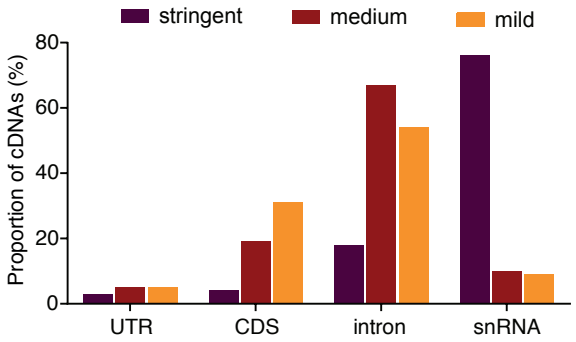
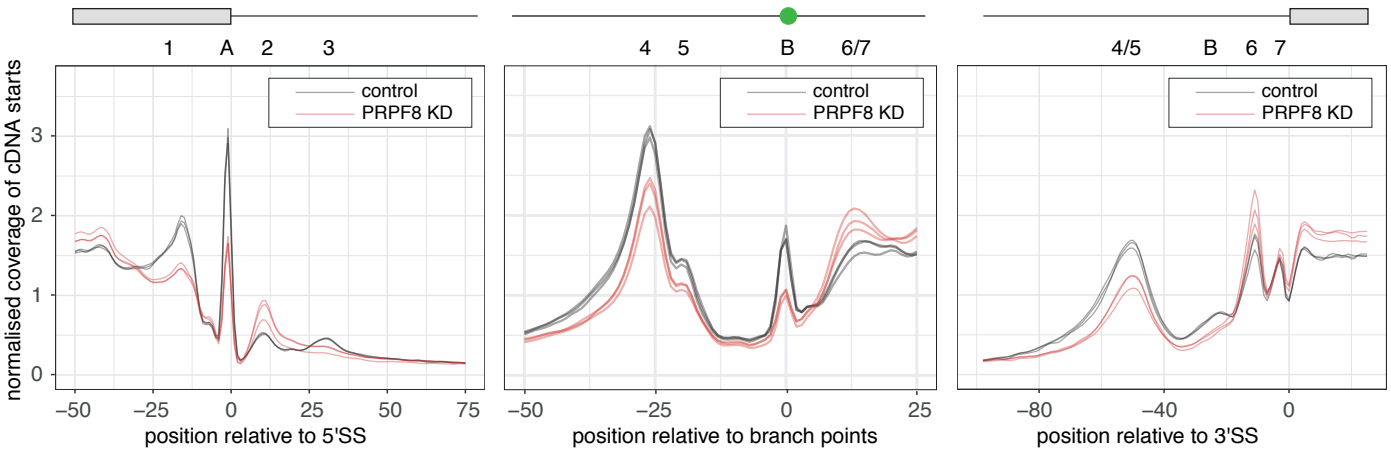
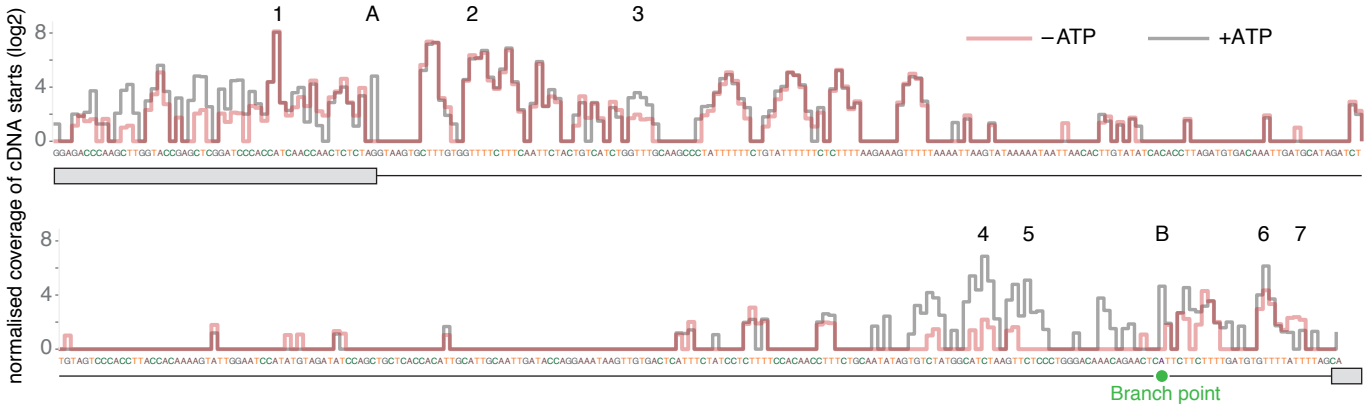

stringent
medium
mild
Proportion of cDNAs (%)
80
60
40
20
0
UTR
CDS
intron
snRNA

Figure 2

A


1
A
2
3
4
5
B
6/7
4/5
B
6
7
control
PRPF8 KD
normalised coverage of cDNA starts
position relative to 5'SS
position relative to branch points
position relative to 3'SS

B


1
A
2
3
4
5
B
6
7
−ATP
+ATP
normalised coverage of cDNA starts (log2)
Branch point

C

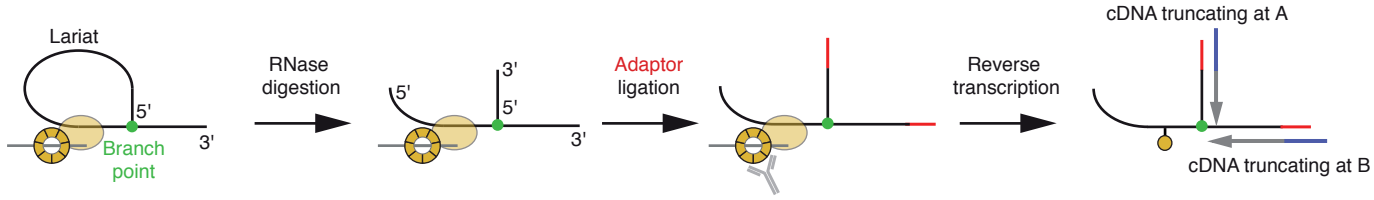
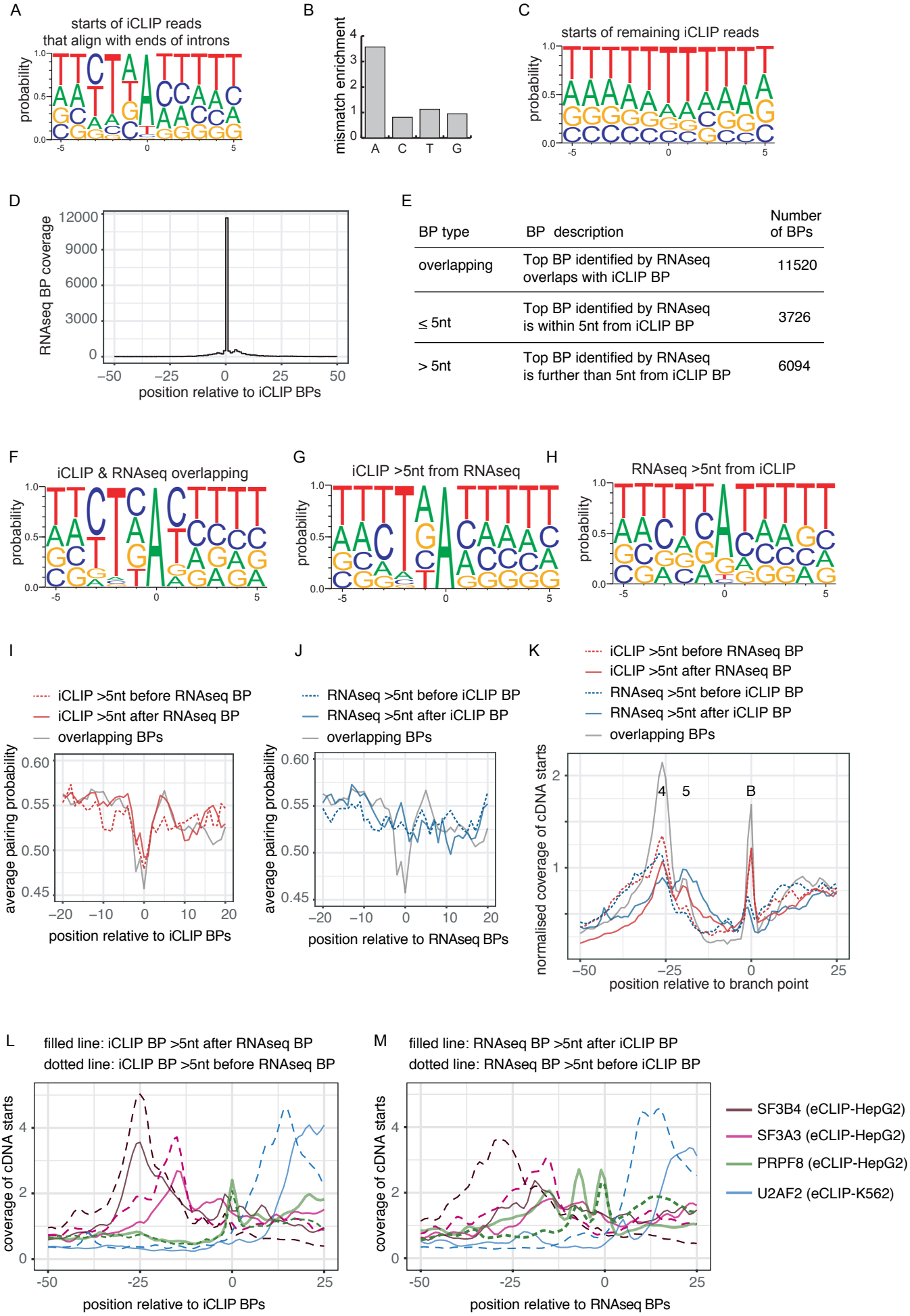

Lariat
5'
3'
Branch point
RNase digestion
Adaptor ligation
Reverse transcription
cDNA truncating at A
cDNA truncating at B

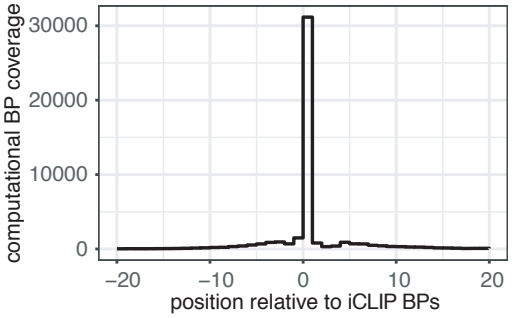
Figure 3



BP type	BP description	Number of BPs
overlapping	Top BP identified by RNAseq overlaps with iCLIP BP	11520
≤ 5nt	Top BP identified by RNAseq is within 5nt from iCLIP BP	3726
> 5nt	Top BP identified by RNAseq is further than 5nt from iCLIP BP	6094

Figure 4

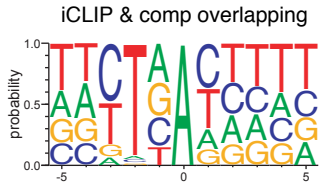
A



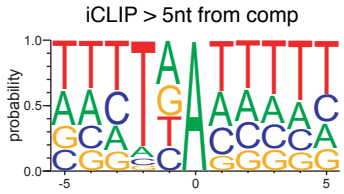
B

BP type	BP description	Number of BPs
overlapping	Top-scoring computational BP overlaps with iCLIP BP	31167
≤ 5nt	Top-scoring computational BP is within 5nt from iCLIP BP	7787
> 5nt	Top-scoring computational BP is further than 5nt from iCLIP BP	11858

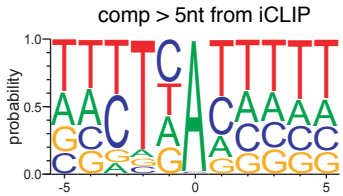
C



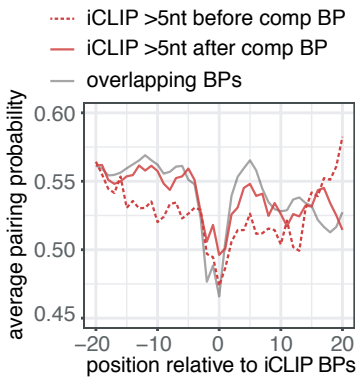
D



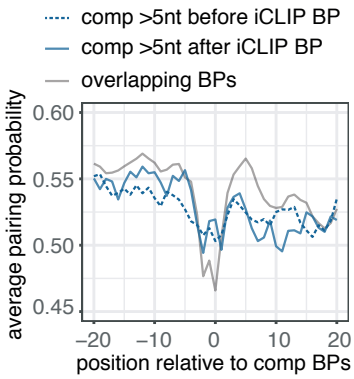
E



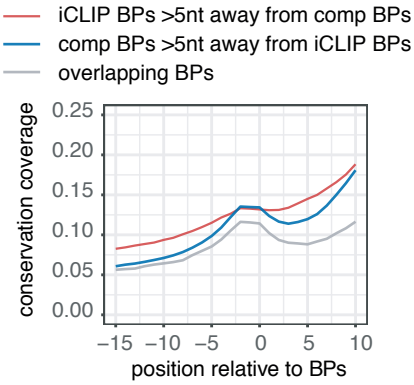
F



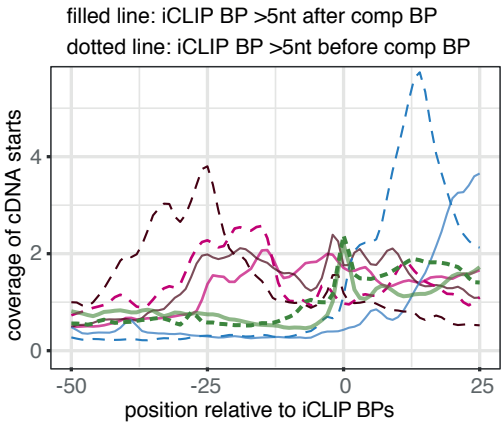
G



H



I



J

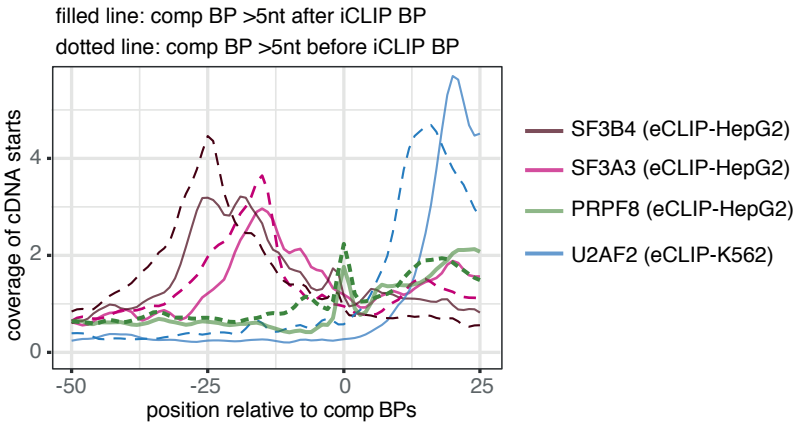


Figure 5

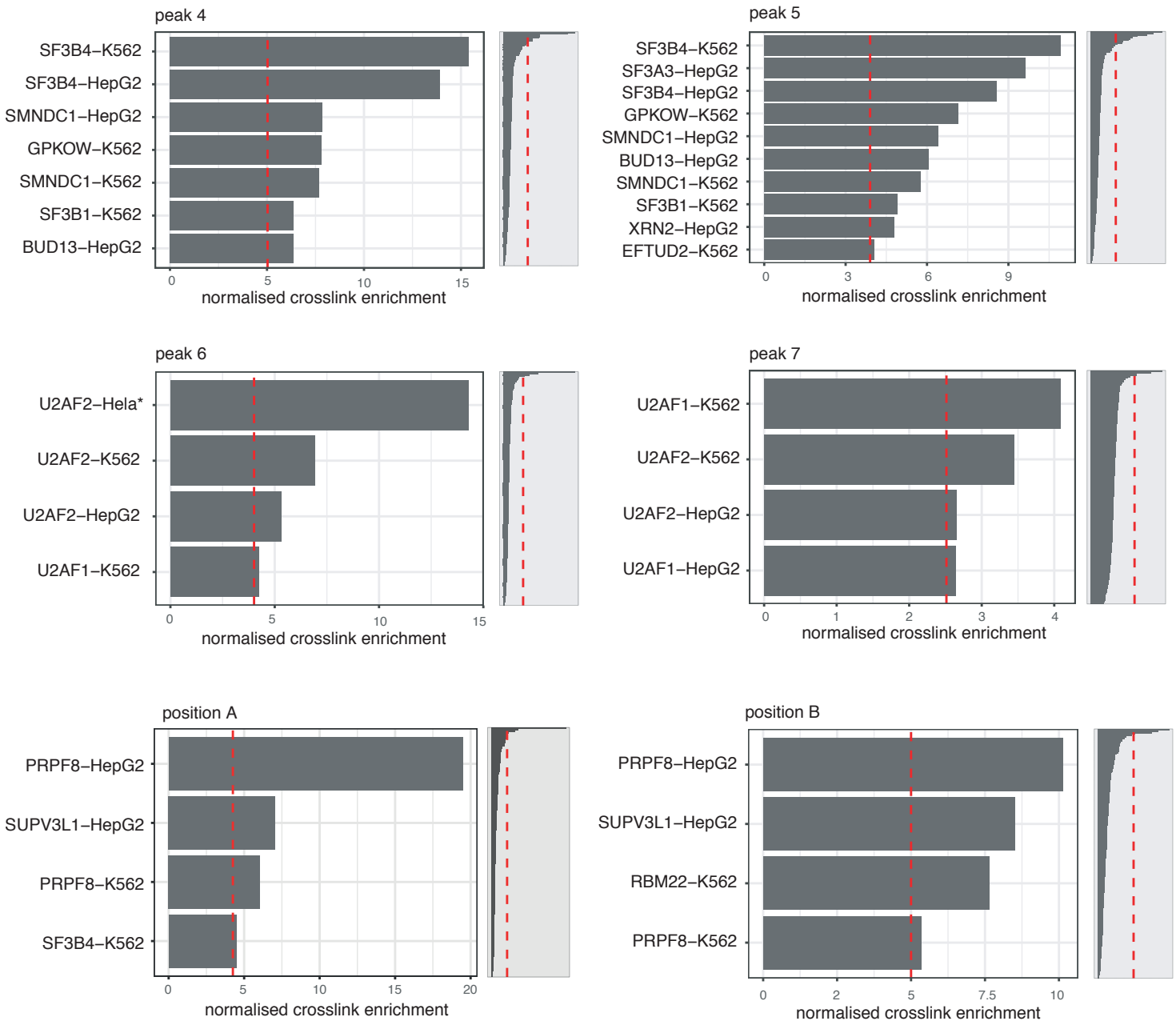

peak 4
SF3B4−K562
SF3B4−HepG2
SMNDC1−HepG2
GPKOW−K562
SMNDC1−K562
SF3B1−K562
BUD13−HepG2
0
5
10
15
normalised crosslink enrichment
peak 5
SF3B4−K562
SF3A3−HepG2
SF3B4−HepG2
GPKOW−K562
SMNDC1−HepG2
BUD13−HepG2
SMNDC1−K562
SF3B1−K562
XRN2−HepG2
EFTUD2−K562
0
3
6
9
normalised crosslink enrichment
peak 6
U2AF2−Hela*
U2AF2−K562
U2AF2−HepG2
U2AF1−K562
0
5
10
15
normalised crosslink enrichment
peak 7
U2AF1−K562
U2AF2−K562
U2AF2−HepG2
U2AF1−HepG2
0
1
2
3
4
normalised crosslink enrichment
position A
PRPF8−HepG2
SUPV3L1−HepG2
PRPF8−K562
SF3B4−K562
0
5
10
15
20
normalised crosslink enrichment
position B
PRPF8−HepG2
SUPV3L1−HepG2
RBM22−K562
PRPF8−K562
0
2
5
7.5
10
normalised crosslink enrichment

Figure 6

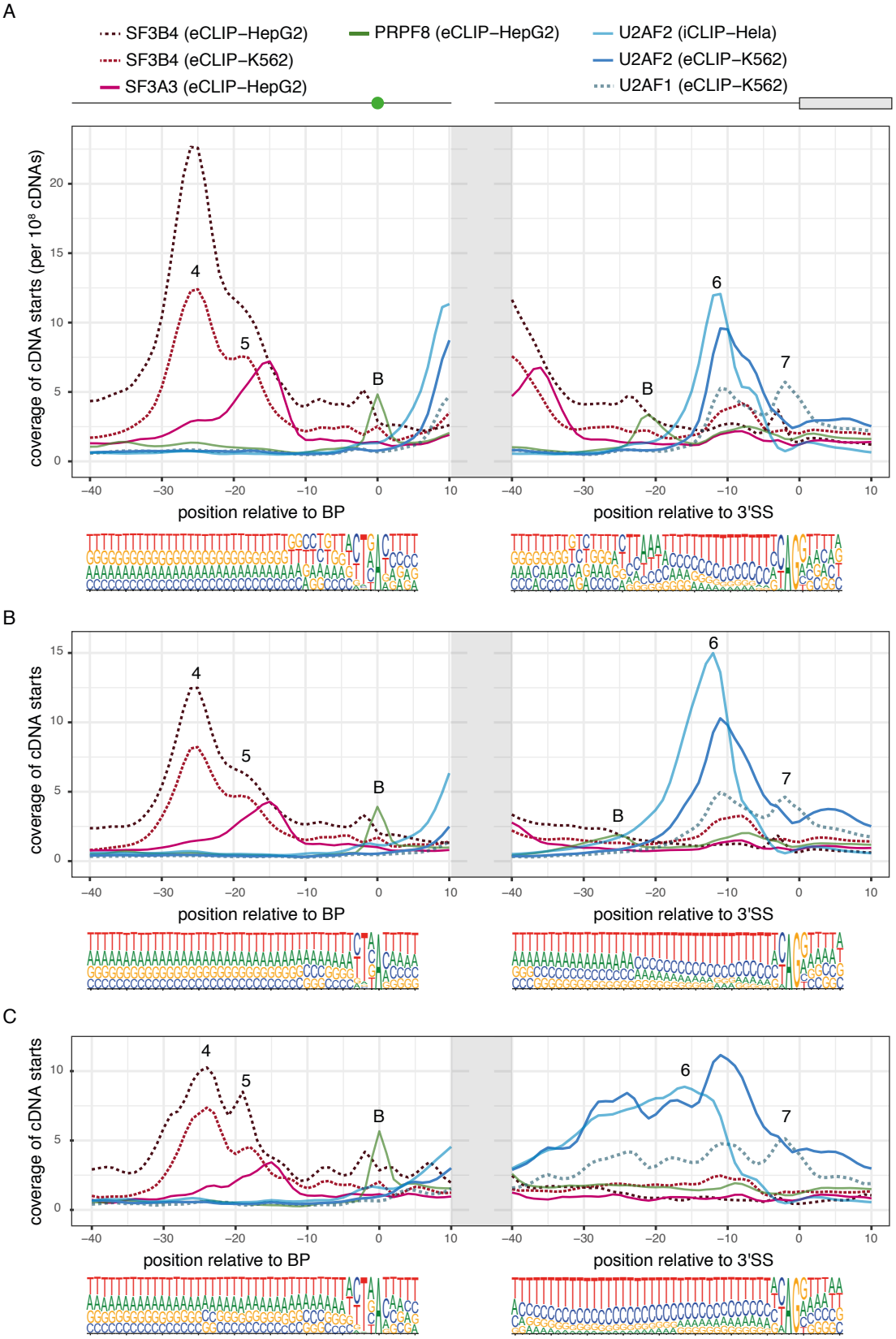

A
SF3B4 (eCLIP−HepG2)
SF3B4 (eCLIP−K562)
SF3A3 (eCLIP−HepG2)
PRPF8 (eCLIP−HepG2)
U2AF2 (iCLIP−Hela)
U2AF2 (eCLIP−K562)
U2AF1 (eCLIP−K562)
coverage of cDNA starts (per 10⁶ cDNAs)
position relative to BP
position relative to 3'SS
B
coverage of cDNA starts
position relative to BP
position relative to 3'SS
C
coverage of cDNA starts
position relative to BP
position relative to 3'SS

Figure 7

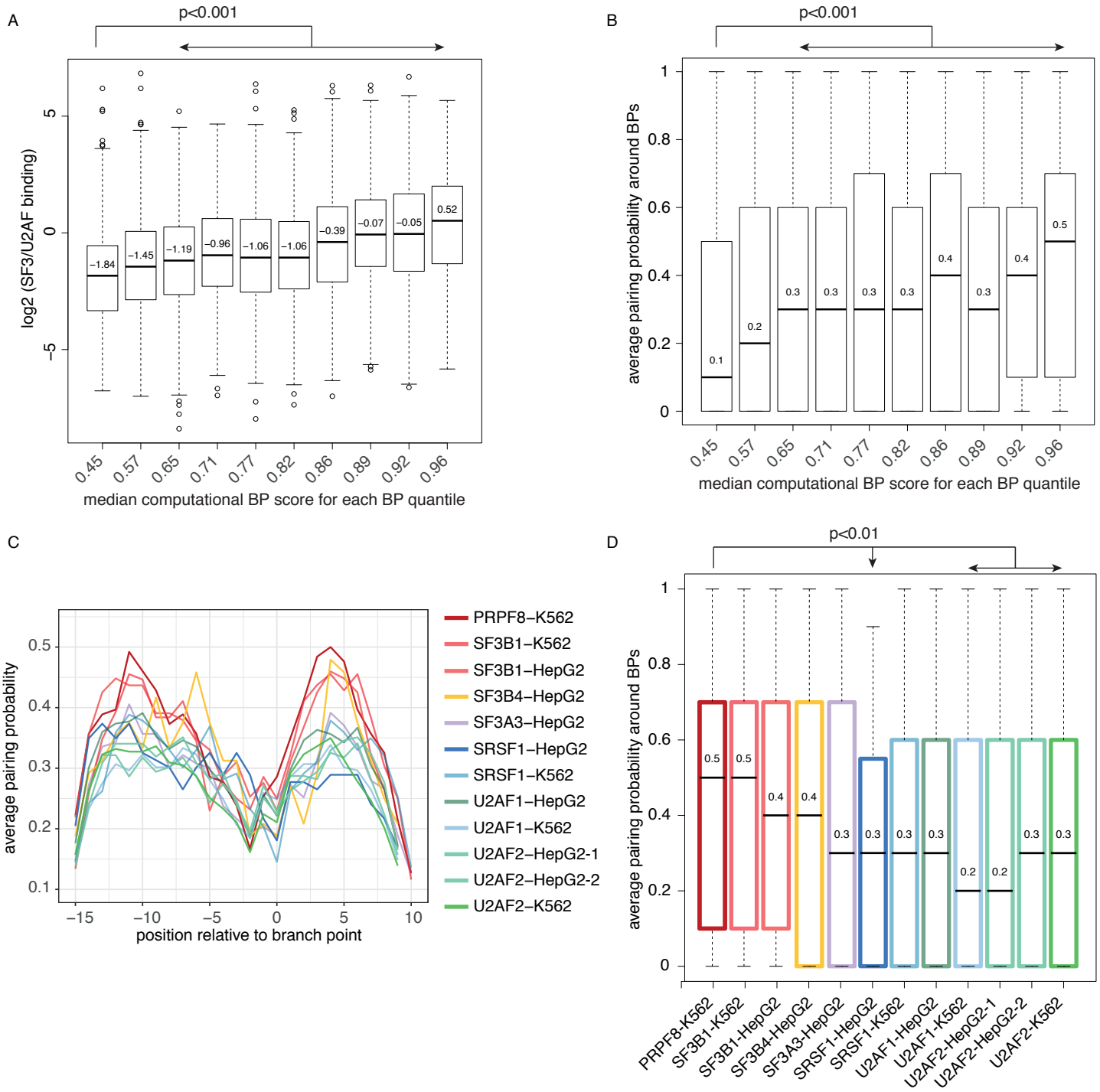

A
p<0.001
log2 (SF3/U2AF binding)
5
0
−5
−1.84
−1.45
−1.19
−0.96
−1.06
−1.06
−0.39
−0.07
−0.05
0.52
0.45
0.57
0.65
0.71
0.77
0.82
0.86
0.89
0.92
0.96
median computational BP score for each BP quantile
B
p<0.001
average pairing probability around BPs
1
0.8
0.6
0.4
0.2
0
0.1
0.2
0.3
0.3
0.3
0.3
0.4
0.3
0.4
0.5
0.45
0.57
0.65
0.71
0.77
0.82
0.86
0.89
0.92
0.96
median computational BP score for each BP quantile
C
average pairing probability
0.5
0.4
0.3
0.2
0.1
−15
−10
−5
0
5
10
position relative to branch point
PRPF8−K562
SF3B1−K562
SF3B1−HepG2
SF3B4−HepG2
SF3A3−HepG2
SRSF1−HepG2
SRSF1−K562
U2AF1−HepG2
U2AF1−K562
U2AF2−HepG2-1
U2AF2−HepG2-2
U2AF2−K562
D
p<0.01
average pairing probability around BPs
1
0.8
0.6
0.4
0.2
0
0.5
0.5
0.4
0.4
0.3
0.3
0.3
0.3
0.2
0.2
0.3
0.3
PRPF8-K562
SF3B1-K562
SF3B1-HepG2
SF3B4-HepG2
SF3A3-HepG2
SRSF1-HepG2
SRSF1-K562
U2AF1-HepG2
U2AF1-K562
U2AF2-HepG2-1
U2AF2-HepG2-2
U2AF2-K562

Figure S1. Related to Figure 1

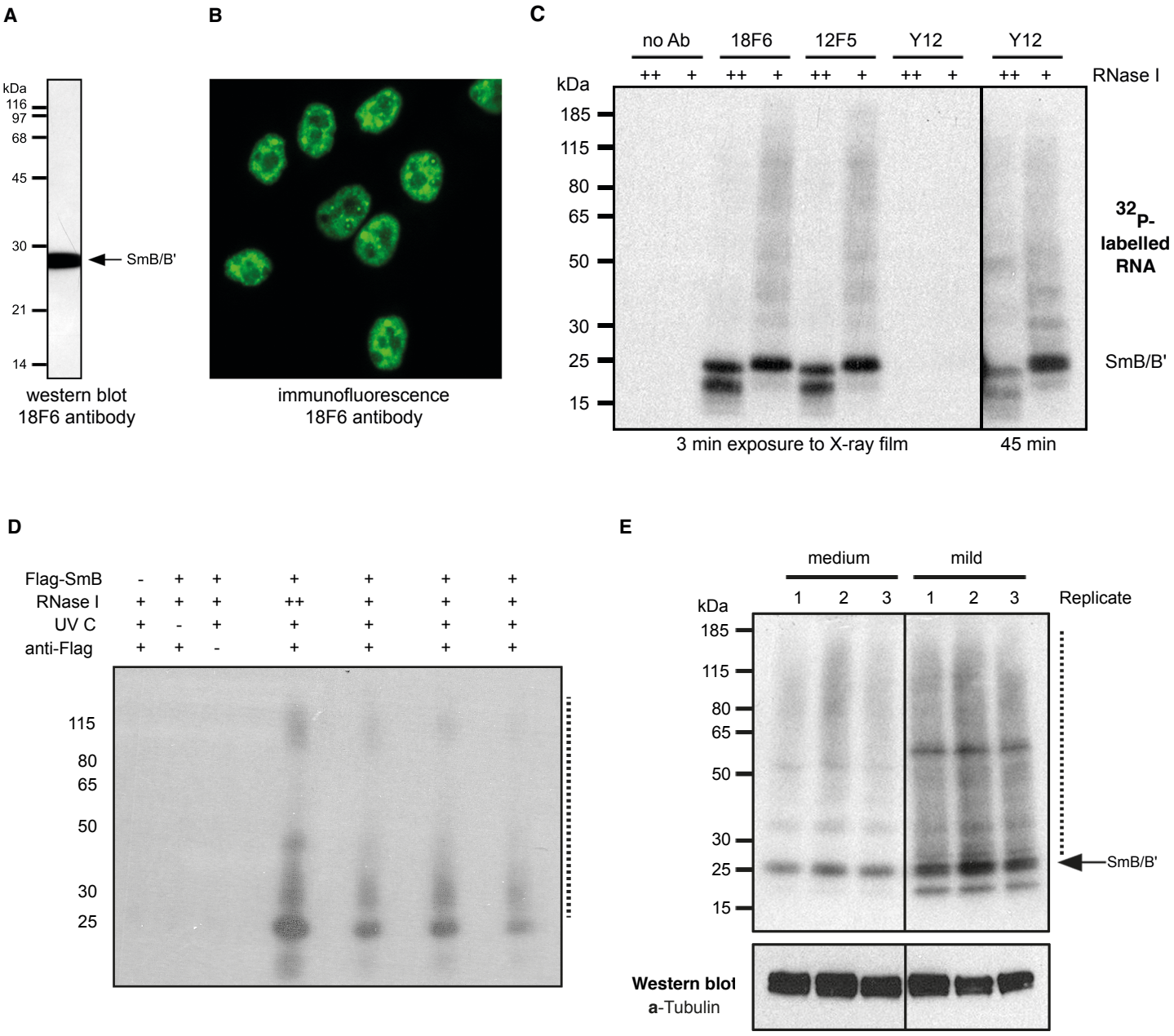
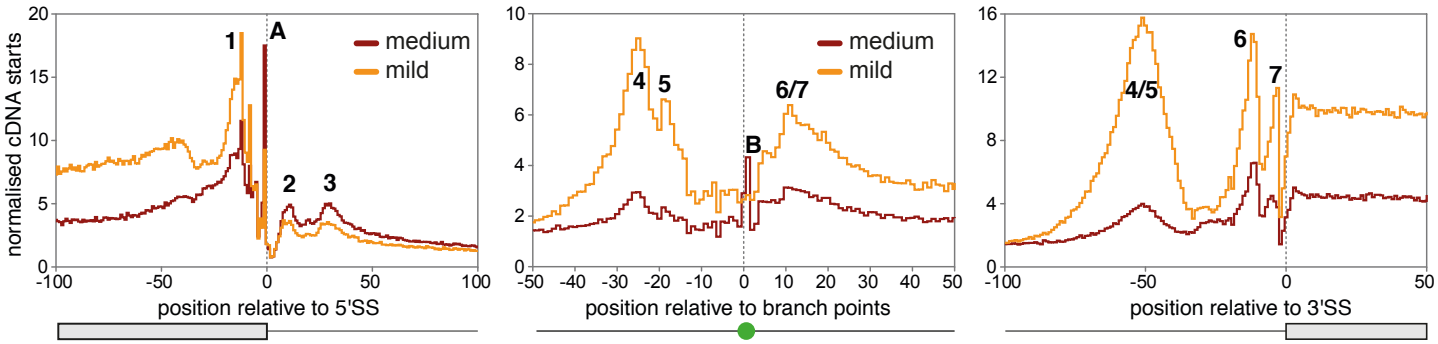
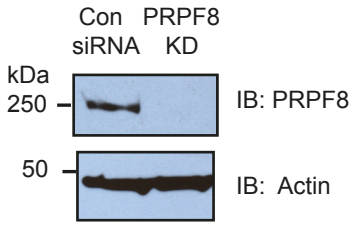


Figure S2. Related to Figure 2

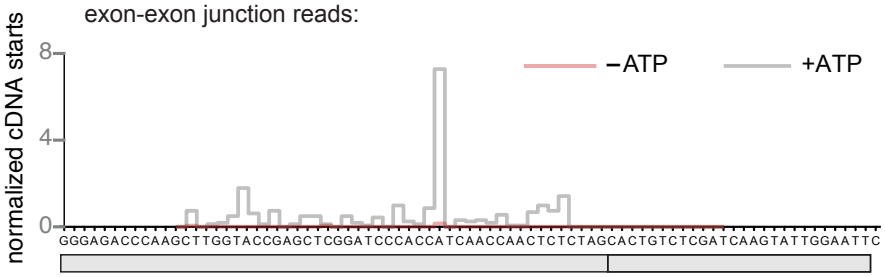
A



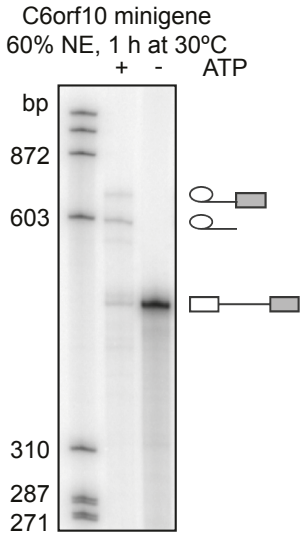
B



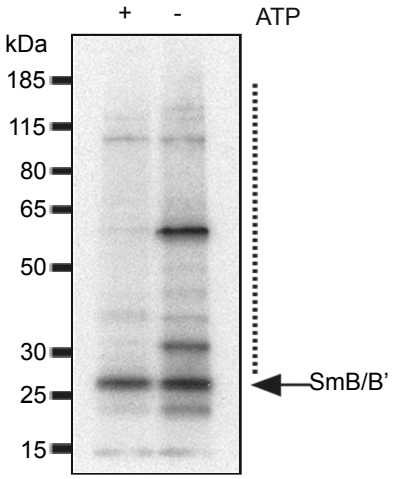
E



C



D



F

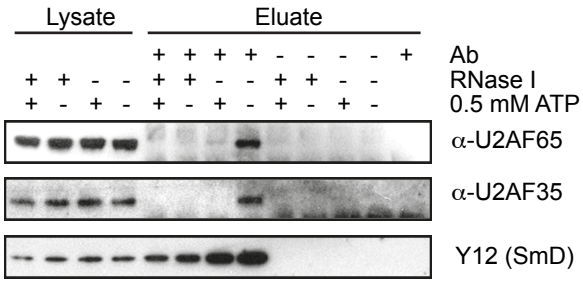
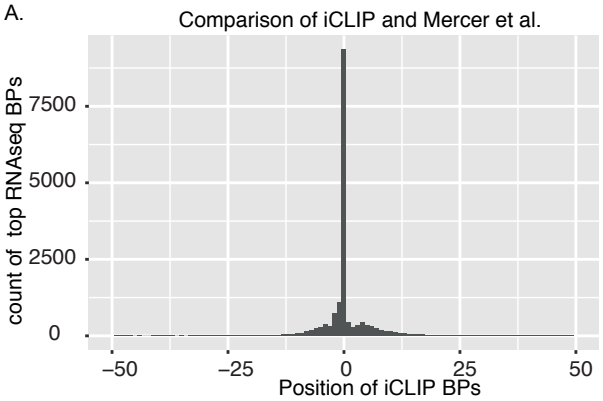
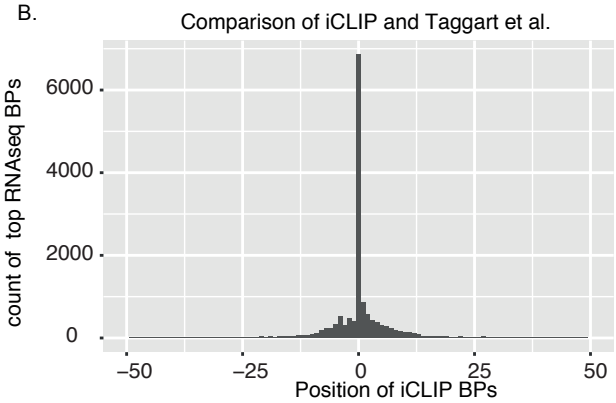


Figure S3. Related to Figure 3

A.



B.

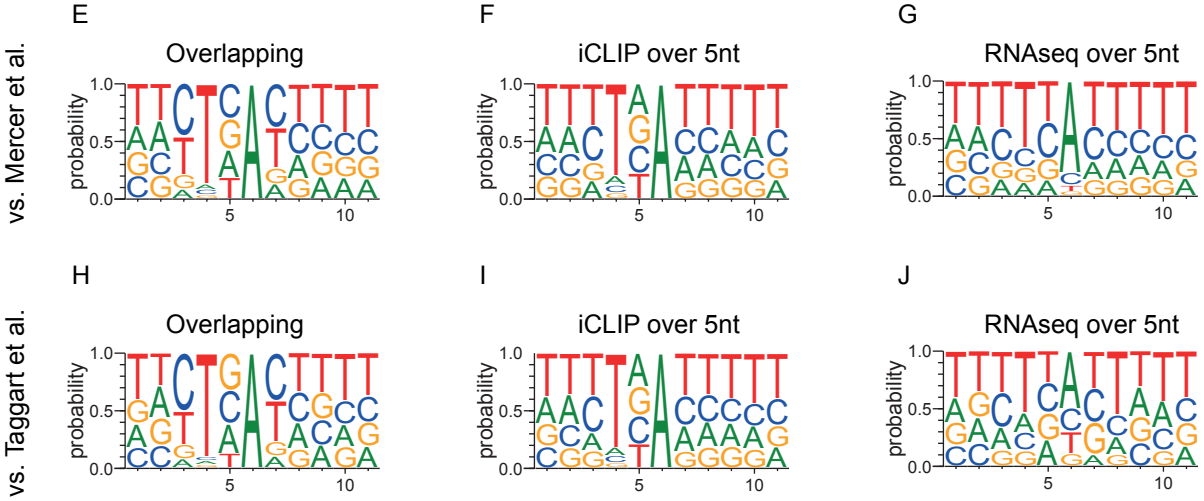


C.

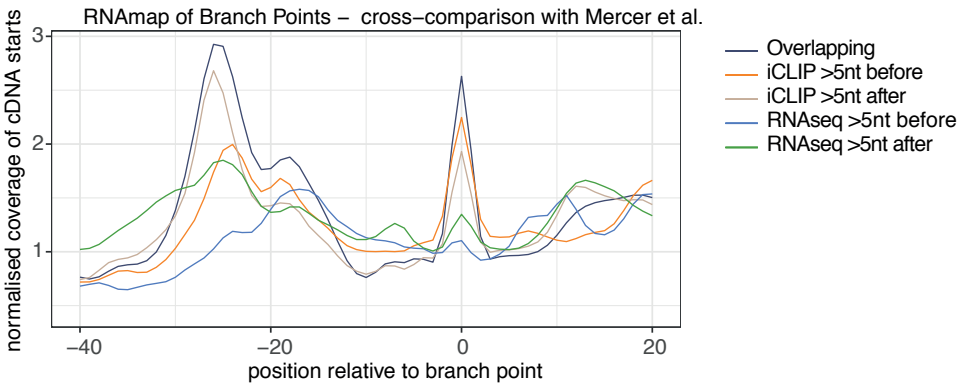
Study	BP positioning	Introns
Mercer et al.	Overlapping with iCLIP BP	9348
	Within 5nt of iCLIP BP	4604
	Over 5nt of iCLIP BP	3095

D.

Study	BP positioning	Introns
Taggart et al.	Overlapping with iCLIP BP	6853
	Within 5nt of iCLIP BP	4553
	Over 5nt of iCLIP BP	3766



K



L

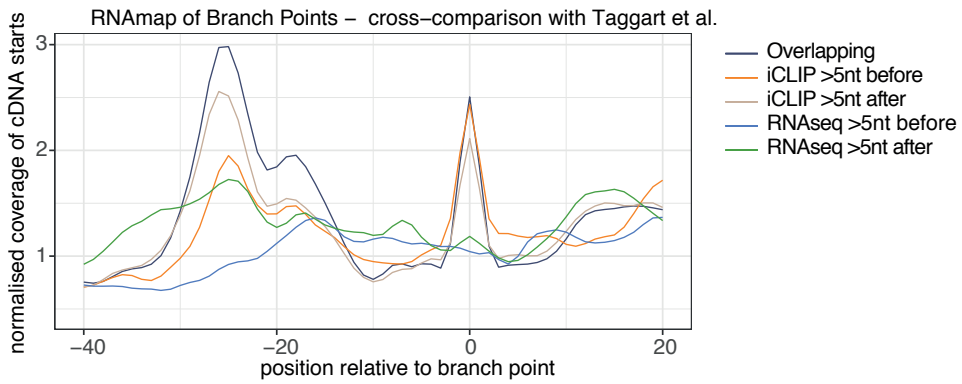
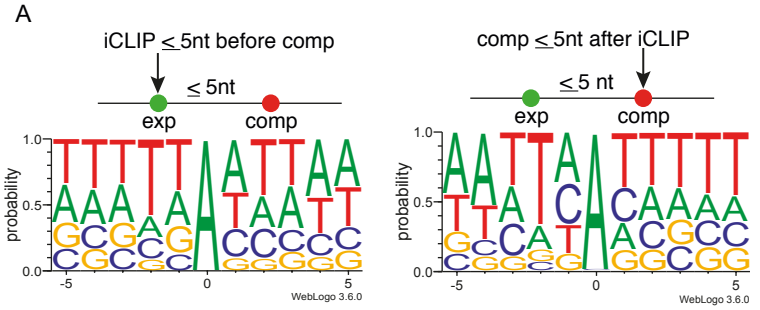
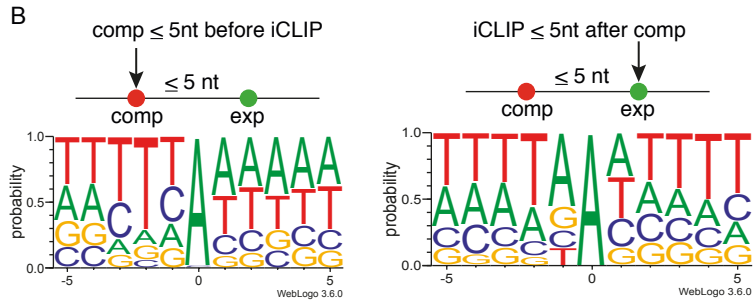


Figure S4. Related to Figure 4

A



B



C

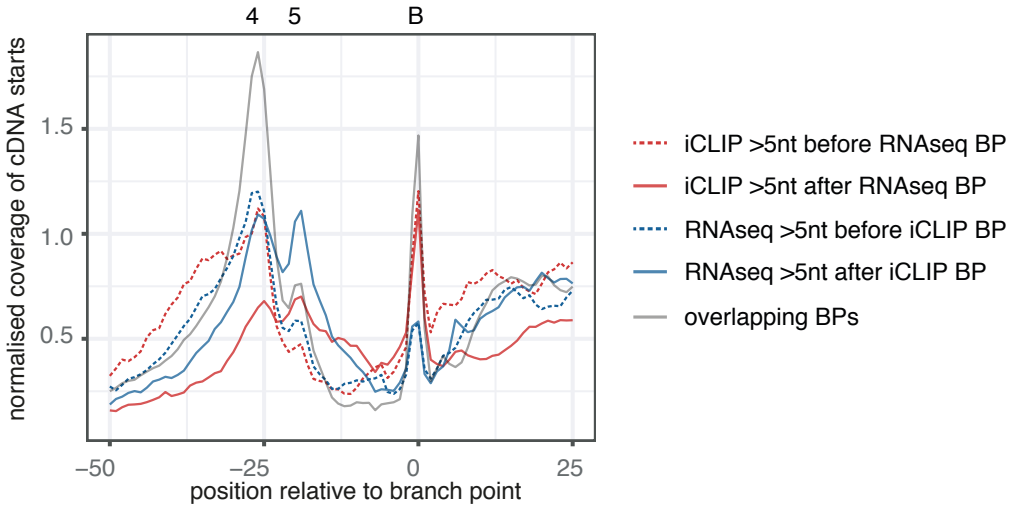
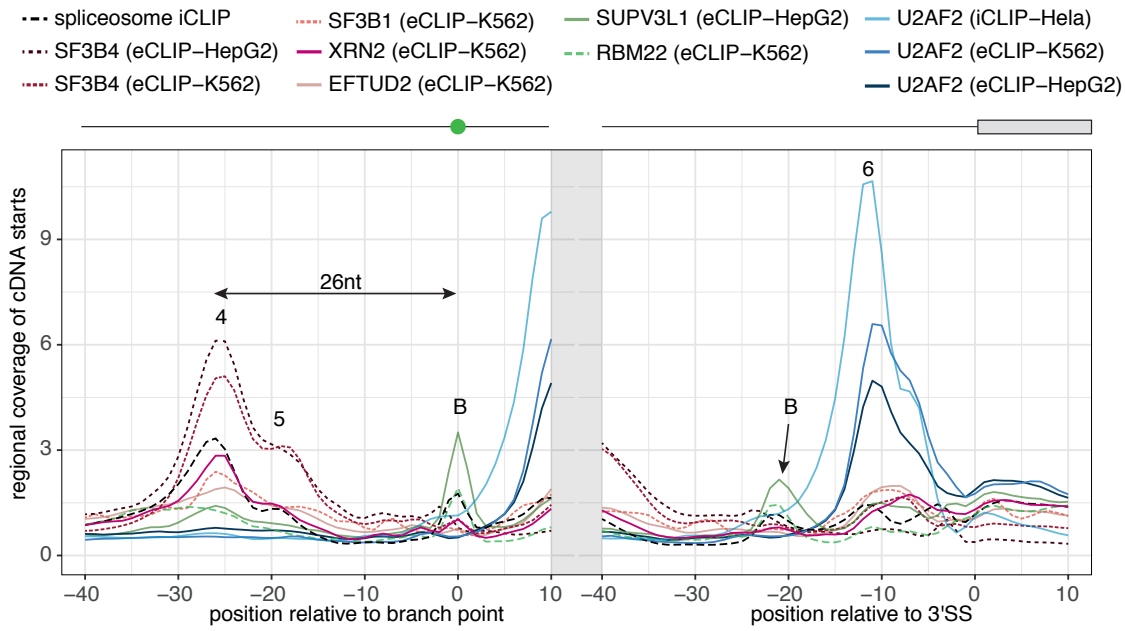
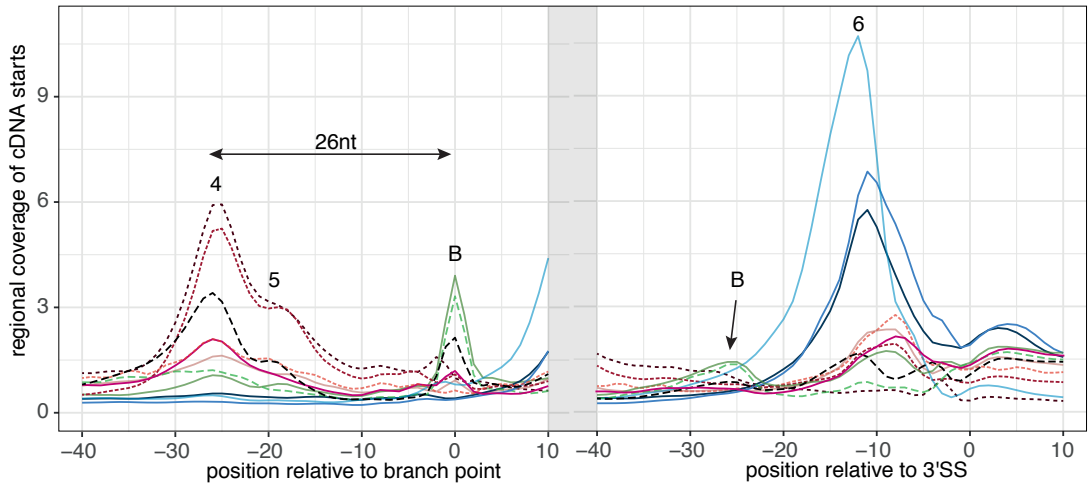


Figure S5. Related to Figure 5

A



B



C

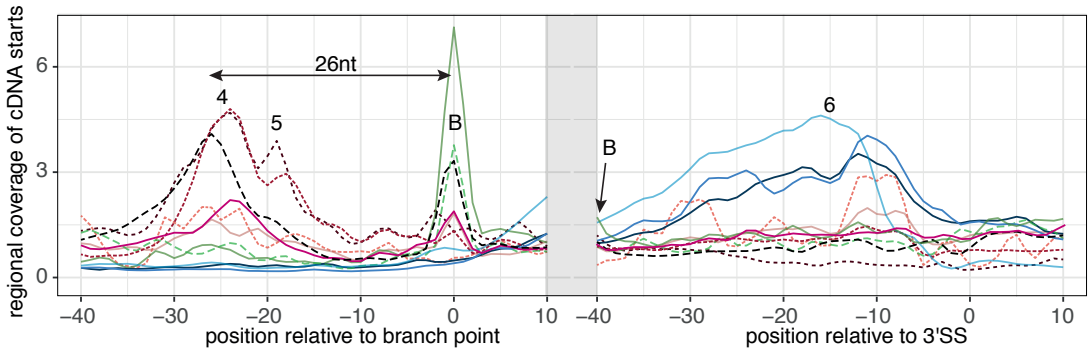


Figure S6. Related to Figure 5

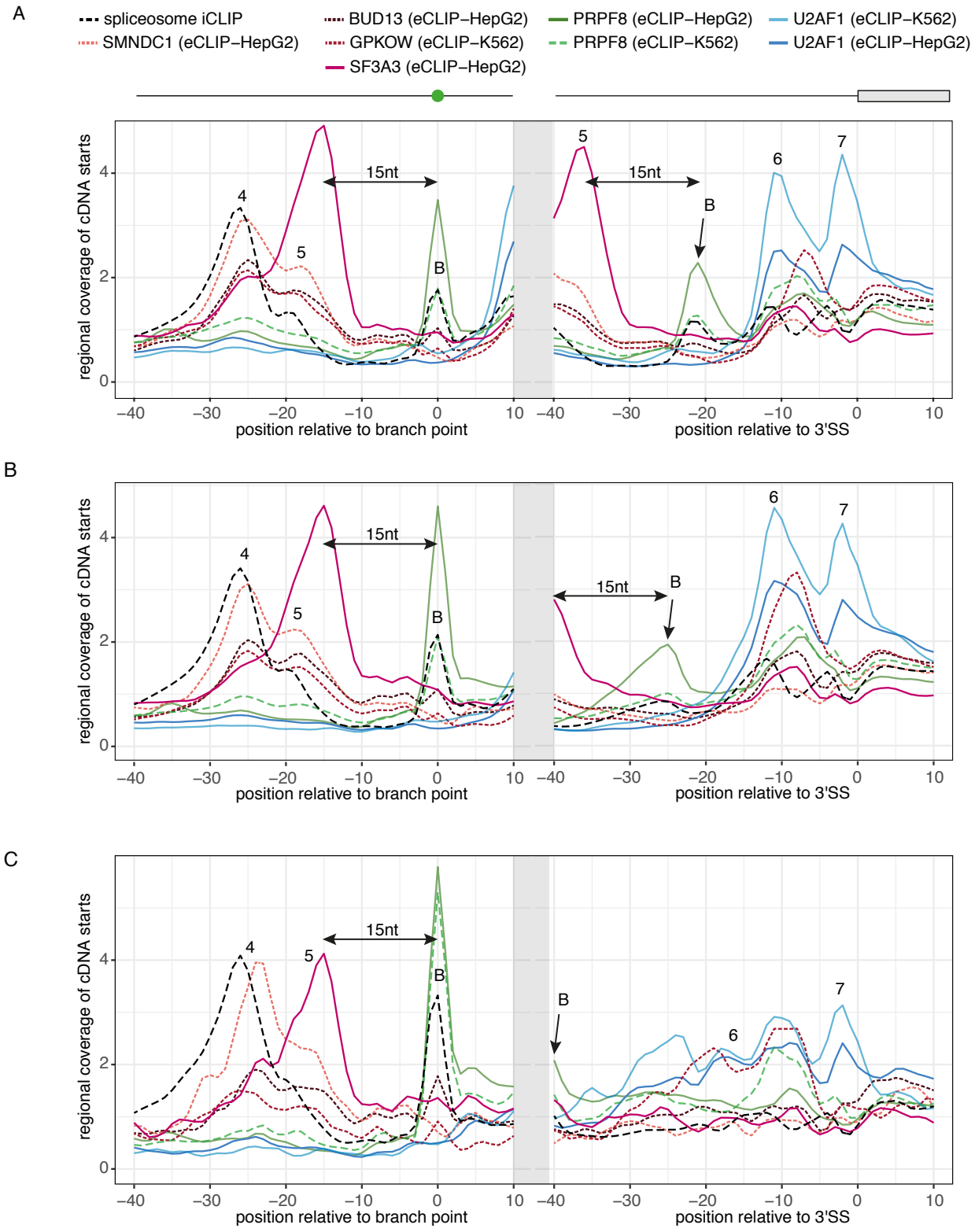


Figure S7. Related to Figure 7

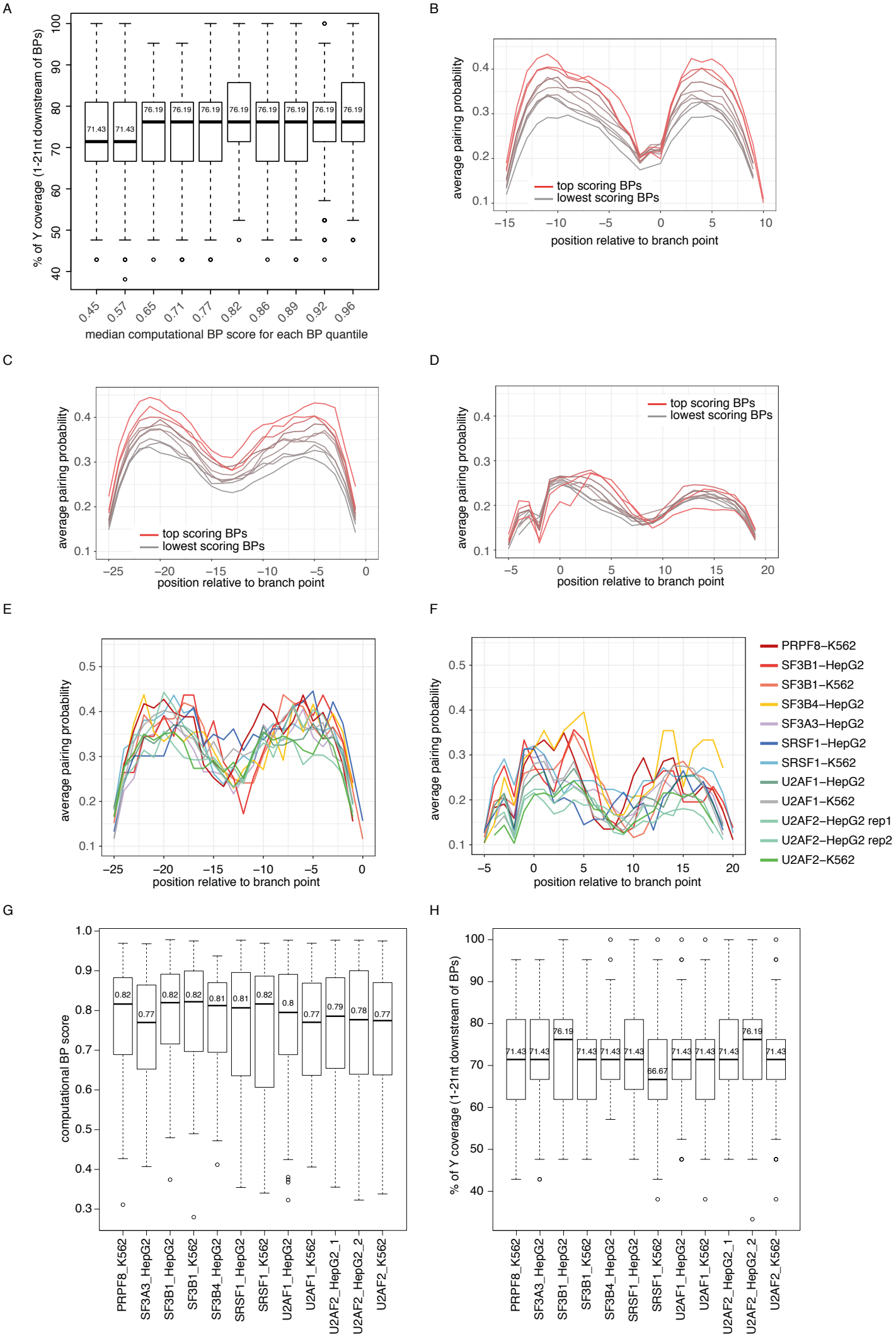


Table S1. Related to Figure 1. iCLIP lysis and wash buffer compositions.

	Lysis buffer			IP Buffer	Wash buffer		
	Stringent	Medium	Mild	Stringent	Stringent	Medium	Mild
Tris-HCl pH 7.4 in mM	50	50	50	50	50	50	50
Urea in M	6	-	-	-	-	-	-
PBS in %	25	-	-	-	-	-	-
NaCl in M	-	0.1	0.1	0.15	1	1	0.1
Igepal CA-630 in %	-	1	0.1	-	1	1	0.1
SDS in %	1	0.1	0.01	-	0.1	0.1	0.01
Na-Deoxycholate in %	-	0.5	0.05	-	0.5	0.5	0.05
Tween 20 in %	-	-	-	0.5	-	-	-
EDTA in mM	-	-	-	0.1	1	1	1

Table S2. Related to Figure 1 and 2. Reverse transcription primer sequences.

Protein	Condition	Sample	Replicate	Primer
SmB/B'	Medium	Mouse brain	1	5'-X33NNAACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			2	5'-X33NNACAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			3	5'-X33NNATTGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
SmB/B'	Mild	Mouse brain	1	5'-X33NNGACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			2	5'-X33NNGTGGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			3	5'-X33NNTATTNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
3xFlag-SmB	Stringent	HEK-293 FLP-in	1	5'-X33NNAACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			2	5'-X33NNACAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			3	5'-X33NNATTGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
SmB/B'	Medium	Cal51, S-phase	1	5'-X33NNGTGGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			2	5'-X33NNTATTNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			3	5'-X33NNTTAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
SmB/B'	Medium	Cal51, G1-phase	1	5'-X33NNAACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			2	5'-X33NNACAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			3	5'-X33NNATTGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
SmB/B'	Medium	Cal51, G2-phase	1	5'-X33NNGTGGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			2	5'-X33NNTTAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			3	5'-X33NNACAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
SmB/B'	Mild	Cal51, control siRNA	1	5'-X33NNAACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
				5'-X33NNCGCCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			2	5'-X33NNACAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
				5'-X33NNGCCANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			3	5'-X33NNATTGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
				5'-X33NNGACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
SmB/B'	Mild	Cal51, PRPF8 siRNA	1	5'-X33NNAACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
				5'-X33NNCGCCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			2	5'-X33NNACAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
				5'-X33NNGCCANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
			3	5'-X33NNATTGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
				5'-X33NNGACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
SmB/B'	Mild	C6orf10, +ATP, HeLa nuclear extract	1	5'-X33NNGACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'
SmB/B'	Mild	C6orf10, -ATP, HeLa nuclear extract	1	5'-X33NNGTGGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC-3'

Table S3. Related to Figure 1 and 2. Number of Sm iCLIP sequence reads mapping to the genome.

Protein	Condition	Sample	Replicate	# Reads matched with barcode	# Reads mapped as single hits to the genome (%)	# Remaining reads mapped as single hits to the transcriptome (%)	# Remaining reads mapped as multiple hits to the transcriptome (%)
SmB/B'	Medium	Mouse brain	1	4113889	1998460 (48.6)	36148 (0.9)	1382606 (33.6)
			2	2066372	1100421 (53.3)	18156 (0.9)	683290 (33.1)
			3	2298511	1179531 (51.3)	19889 (0.9)	756624 (32.9)
SmB/B'	Mild	Mouse brain	1	3762352	1768447 (47)	43900 (1.2)	1215167 (32.3)
			2	3997243	1801292 (45.1)	49729 (1.2)	1261197 (31.6)
			3	878795	438603 (49.9)	11167 (1.3)	267526 (30.4)
SmB	Stringent	HEK-293 FLP-in	1	1863149	831903 (44.7)	187158 (10)	26770 (1.4)
			2	7647	1535 (20.1)	346 (4.5)	43 (0.6)
			3	537925	247203 (46)	56425 (10.5)	9180 (1.7)
SmB/B'	Medium	Cal51, S-phase	1	4229497	2376499 (56.2)	1063010 (25.1)	57707 (1.4)
			2	3487327	2265756 (65)	562253 (16.1)	67754 (1.9)
			3	33808485	21734025 (64.3)	5266589 (15.6)	670695 (2)
SmB/B'	Medium	Cal51, G1-phase	1	2752581	1464301 (53.2)	785779 (28.5)	33280 (1.2)
			2	3611115	2267936 (62.8)	648596 (18.0)	68367 (1.9)
			3	3499621	2121357 (60.6)	679628 (19.4)	67301 (1.9)
SmB/B'	Medium	Cal51, G2-phase	1	5771108	2742102 (47.5)	1713027 (29.7)	149164 (2.6)
			2	3243956	1740684 (53.7)	790046 (24.4)	77943 (2.4)
			3	2836975	1429392 (50.4)	764979 (27)	62173 (2.2)
SmB/B'	Mild	Cal51, control siRNA	1	20710253	10790305 (52.1)	3960442 (19.1)	1187137 (5.7)
			2	125806511	66258207 (52.7)	19837136 (15.8)	7168086 (5.7)
			3	106102741	53149959 (50.1)	18861424 (17.8)	6795511 (6.4)
SmB/B'	Mild	Cal51, PRPF8 siRNA	1	61948551	29208450 (47.1)	15105468 (24.4)	3263369 (5.3)
			2	163700256	85409913 (52.2)	30054667 (18.4)	10174864 (6.2)
			3	62329446	32514674 (52.2)	11947025 (19.2)	3194568 (5.1)

Table S4. Related to Figure 2. Number of Sm iCLIP reads mapping to C6orf10 unspliced and spliced substrates.

Protein	Condition	Sample	Replicate	# Reads matched with barcode	# Reads processed for mapping	# Reads mapping to the unspliced substrate	# Reads mapping to the spliced substrate
SmB/B'	Mild	C6orf10, +ATP, HeLa nuclear extract	1	4417628	4409507	48584	364
SmB/B'	Mild	C6orf10, -ATP, HeLa nuclear extract	1	3515991	3506903	45277	5