A systems view of spliceosomal assembly and branchpoints with iCLIP

Michael Briese[1,2]*, Nejc Haberman[3,4]*, Christopher R. Sibley[1,4,5,6]*, Rupert Faraway[3,4], Andrea S. Elser[3,4], Anob M. Chakrabarti[3,7], Zhen Wang[1], Julian König[1,8], David Perera[9], Vihandha O. Wickramasinghe[9,10], Ashok R. Venkitaraman[9], Nicholas M. Luscombe[3,7,11], Luciano Saieva[12,13], Livio Pellizzoni[12], Christopher W.J. Smith[14], Tomaž Curk[15], Jernej Ule[1,3,4]§

[1]MRC Laboratory of Molecular Biology, Cambridge, UK

[2]Institute of Clinical Neurobiology, University of Wuerzburg, Wuerzburg, Germany

[3]The Francis Crick Institute, London, UK

[4]Department of Neuromuscular Disease, UCL Institute of Neurology, London, UK

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

[6]Institute of Quantitative Biology, Biochemistry and Biotechnology, Edinburgh University, UK

[7]Department of Genetics, Environment and Evolution, UCL Genetics Institute, London, UK

[8]Institute of Molecular Biology (IMB) GmbH, Mainz, Germany

[9]MRC Cancer Unit at the University of Cambridge, Cambridge, UK

[10]RNA Biology and Cancer Laboratory, Peter MacCallum Cancer Centre, Melbourne, Australia

[11]Okinawa Institute of Science & Technology Graduate University, Okinawa, Japan

[12]Center for Motor Neuron Biology and Disease, Department of Pathology and Cell Biology, Columbia University, New York, NY, USA

[13]Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK

[14]Department of Biochemistry, University of Cambridge, Cambridge, UK

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

Equal contributions:
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 map spliceosome engagement with pre-mRNAs in human cell lines. This revealed seven peaks of spliceosomal crosslinking around branchpoints (BPs) and splice sites. We identified RBPs that crosslink to each peak, including known and candidate splicing factors. Moreover, we detected use of over 40,000 BPs with strong sequence consensus and structural accessibility, which align well to nearby crosslinking peaks. We show how the position and strength of BPs affect the crosslinking patterns of spliceosomal factors, which bind more efficiently upstream of strong or proximally located BPs, and downstream of weak or distally located BPs. These insights exemplify spliceosome iCLIP as a broadly applicable method for transcriptomic studies of splicing mechanisms.

Introduction

Splicing is a multi-step process in which small nuclear ribonucleoprotein particles (snRNPs) and associated splicing factors bind at specific positions around intron boundaries in order to assemble an active spliceosome through a series of remodeling steps. The splicing reactions are coordinated by dynamic pairings between different snRNAs, between snRNAs and pre-mRNA, and by protein-RNA contacts[1]. Spliceosome assembly begins with ATP-independent binding of U1 snRNP at the 5' splice site (ss), and of U2 small nuclear RNA auxiliary factors 1 and 2 (U2AF1 and U2AF2, also known as U2AF35 and U2AF65) to the 3'ss. ATP-dependent remodeling then leads to the formation of complex A in which U2 snRNP contacts the branchpoint (BP), stabilized through interactions with the U2AF and U2 snRNP splicing factor 3 (SF3a and SF3b) complex. Next, U4/U6 and U5 snRNPs are recruited to form complex B. The actions of many RNA helicases and pre-mRNA processing factor 8 (PRPF8) then facilitate rearrangements of snRNP interactions and establishment of the catalytically competent B_{act} and C complexes. These catalyze the two trans-esterification reactions leading to lariat formation, intron removal and exon ligation[2].

Transcriptome-wide studies of splicing reactions are valuable to unravel the multi-component and dynamic assembly of the spliceosome on the pre-mRNA substrate[3-5]. Accordingly, "spliceosome profiling" has been developed through affinity purification of the tagged U2·U5·U6·NTC complex from *Schizosaccharomyces pombe* to monitor its interactions using a RNA footprinting-based strategy[3,4]. However, it is unclear if this method can be applied to mammalian cells which might be more sensitive to introduction of affinity tags into splicing factors. Furthermore, no method has simultaneously monitored the full complexity of the interactions of diverse RBPs on pre-mRNAs from the earliest to the latest stages of spliceosomal assembly.

Here, we have adapted the individual nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) method[6] to develop spliceosome iCLIP. This approach identifies crosslinks of endogenous, untagged spliceosomal factors on pre-mRNAs at nucleotide resolution. In a previous study, we demonstrated validity of this approach by showing how PRPF8 remodels spliceosomal contacts at 5'ss[5]. Here, we comprehensively characterize spliceosome iCLIP and show that it simultaneously maps the crosslink profiles of core and accessory spliceosomal factors that are known to participate across the diverse stages of the splicing cycle. Due to iCLIP's nucleotide precision, we distinguished 7 binding peaks corresponding to distinct RBPs that differ in their requirement for ATP or the factor PRPF8. Spliceosome iCLIP also purifies intron lariats and identified 132,287 candidate BP positions. Compared to BPs identified in previous RNA-seq studies[7-9], those identified by spliceosome iCLIP contain more canonical sequence and structural features. We further examined the binding profiles of spliceosomal RBPs around the BPs. This demonstrates that assembly of SF3 and associated spliceosomal complexes tends to be determined by a primary BP in most introns, even though alternative BPs are detected by lariat-derived reads in RNA-seq. Moreover, we identify complementary roles of U2AF and SF3 complexes in BP definition. Taken together, these findings demonstrate the value of spliceosome iCLIP for transcriptome-wide studies of BP definition and spliceosomal interactions with pre-mRNAs.

Results

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

SmB/B' proteins are part of the highly stable Sm core common to all spliceosomal snRNPs except U6[1]. In order to adapt iCLIP for the study of a multi-component machine like the spliceosome, we immunopurified endogenous SmB/B' proteins[10] using a range of conditions with differing stringency of detergents and salt concentrations for the lysis and washing steps (Supplementary Table 1, Fig. 1a and Supplementary Fig. 1a,b). First, to enable denaturing purification, we generated HEK293 cells stably expressing Flag-tagged SmB and employed 6M urea during cell lysis to minimize co-purification of additional proteins[11] ('stringent' purification, Supplementary Table 1), followed by dilution of the lysis buffer (see Methods) to facilitate immunopurification of SmB via the Flag tag. We observed a 25 kDa band corresponding to the molecular weight of SmB-RNA complexes, which was absent when UV light or anti-Flag antibody were omitted, or when cells not expressing Flag-SmB were used (Supplementary Fig. 1c). Next, we employed the standard, non-denaturing iCLIP condition, which uses a high concentration of detergents in the lysis buffer, and wash buffer with 1M NaCl ('medium' purification, Supplementary Table 1). This disrupts most protein-protein interactions but can preserve stable complexes such as snRNPs, as evident by the multiple radioactive bands in addition to the 25 kDa SmB-RNA complex upon treatment with low RNase (Fig. 1b). Of note, similar profiles of protein-RNA complexes were obtained when using different monoclonal SmB/B' antibodies (Supplementary Fig. 1d). Last, we further decreased the concentration of detergents in the lysis buffer, used 0.1M NaCl in the washing buffer ('mild' purification, Supplementary Table 1), and employed the low RNase treatment that leaves snRNAs generally intact such that they serve as a scaffold for purifying the multi-protein spliceosomal complexes (Fig. 1a).

To produce cDNA libraries with spliceosome iCLIP, we immunoprecipitated SmB/B' under the three different stringency conditions from lysates of UV-crosslinked cells, and isolated a broad size distribution of protein-RNA complexes in order to recover the greatest possible diversity of spliceosomal protein-RNA interactions (Fig. 1b and Supplementary Fig. 1c,d). An antibody against endogenous SmB/B' was used for medium and mild purification from HEK293, K562 and HepG2 cells, and an anti-Flag antibody for stringent purification from HEK293 cells expressing Flag-SmB (Supplementary Table 2 and 3). As in previous iCLIP studies[6], the nucleotide preceding each cDNA was used for all analyses. When stringent conditions were used, >75% of iCLIP cDNAs mapped to snRNAs, likely corresponding to the direct binding of Flag-SmB (Fig. 1c). However, the proportion of snRNA crosslinking reduced to ~40-60% under mild and medium conditions, with a corresponding increase of crosslinking to introns and exons that likely reflects binding of snRNP-associated proteins to pre-mRNAs (Fig. 1a,c).

Spliceosome iCLIP identifies seven crosslinking peaks on pre-mRNAs

Assembly of the spliceosome on pre-mRNA is guided by three main landmarks: the 5'ss, 3'ss and BP. Therefore, we evaluated if spliceosomal crosslinks are located at specific positions relative to splice sites and computationally predicted BPs[12]. For this purpose we performed spliceosome iCLIP from human Cal51 cells, which we have previously used as a model system to study the roles of spliceosomal factors in cell cycle[5]. RNA maps of summarized spliceosomal crosslinking revealed 7 peaks around these landmarks (Fig. 2a). Importantly, similar positional patterns were also seen in HEK293, K562 and HepG2 cell lines (Supplementary Fig. 2a). The centers of the peaks were 15 nt upstream of the 5'ss (peak 1), 10 nt downstream of the 5'ss (peak 2), 31 nt downstream of the 5'ss (peak 3), 26 nt upstream of the BP (peak 4), 20 nt upstream of the BP (peak 5), 11 nt upstream of the 3'ss (peak 6) and 3 nt upstream of the 3'ss (peak 7). We also observed alignment of cDNA starts to the start of the intron and the BPs, which we refer to as positions A and B, respectively (Fig. 2a and Supplementary Fig. 2a). The crosslinking enrichment at most peaks was generally stronger under the mild condition, especially at the 3'ss (Supplementary Fig. 2a). This indicates that spliceosome iCLIP performed under mild conditions is most suitable for investigating spliceosomal assembly on pre-mRNAs.

Spliceosome iCLIP monitors multiple stages of spliceosomal remodeling

Next, we investigated whether spliceosome iCLIP is able to monitor spliceosome assembly at different stages during the splicing cycle. For this purpose we knocked down (KD) PRPF8 in Cal51 cells (Supplementary Fig. 2b) and performed spliceosome iCLIP under mild conditions. As an integral component of the U4/U6.U5 tri-snRNP, PRPF8 is essential for both catalytic reactions[1]. We previously showed that PRPF8 is required for efficient spliceosomal assembly at 5'ss[5]. Here, we additionally find that PRPF8 is essential for efficient spliceosomal assembly at peaks 4 and 5 (Fig. 2a). Moreover, we also observed a major decrease of reads truncating at the positions A and B, whereas crosslinking at peaks 2 and 6 is increased upon PRPF8 KD.

To further investigate whether spliceosome iCLIP can monitor distinct stages of the splicing reaction**,** we performed an *in vitro* splicing assay in which an exogenous pre-mRNA splicing substrate was incubated with HeLa nuclear extract in the presence or absence of ATP. ATP is required for the progression of early, ATP-independent, spliceosomal complexes to later assembly stages mediating the catalytic splicing reactions. The RNA substrate was produced by *in vitro* transcription of a minigene construct containing a short intron and flanking exons from the human *C6orf10* gene. Gel electrophoresis analysis confirmed that the minigene RNA was efficiently spliced *in vitro* in an ATP-dependent manner (Supplementary Fig. 2c). We performed spliceosome iCLIP from the splicing reactions using the mild purification condition (Supplementary Fig. 2d). Following sequencing, the reads mapping to the exogenous splicing substrate or spliced product represented ~1%, whereas the remaining reads were derived from endogenous RNAs present in the nuclear extract (Supplementary Table 4). The spliced product was detected with exon-exon junction reads primarily in the presence of ATP (364 reads in +ATP vs. 5 reads in -ATP condition) (Supplementary Fig. 2e and Supplementary Table 4).

As expected given that the spliceosome rapidly disassembles upon completion of the splicing reaction, very few reads mapped to the spliced (364 reads) compared to unspliced substrate (48,584 reads) (Supplementary Table 4) in the +ATP condition. It should be considered, however, that some reads from exogenous minigene could represent RNA that did not enter the splicing pathway.

We visualized crosslinking on the substrate RNA, and marked positions that correspond to peaks on the transcriptome-wide RNA maps (Fig. 2b). Whilst crosslinking peaks on a metagene plot might not necessarily be representative of individual splicing substrates, we nevertheless observed crosslinking in corresponding regions of the *C6orf10* substrate (comparing Fig. 2a and 2b). When comparing crosslinking in the presence or absence of ATP, an unchanged crosslinking profile was seen in regions of peaks 1, 2, 6 and 7, indicating these are ATP-independent contacts of early spliceosomal factors. In contrast, the presence of ATP led to a ~11 fold increase of crosslinking in the region upstream of the BP where the PRPF8-dependent peaks 4 and 5 are located on endogenous transcripts (Fig. 2b). This indicates that spliceosome iCLIP detects pre-mRNA binding of factors contributing to early, ATP-independent and late, ATP-dependent stages of spliceosomal assembly.

Following crosslinking, the peptide that remains bound to the RNA after RBP digestion will normally terminate reverse transcription to produce so-called 'truncated cDNAs'[13-15]. Accordingly, analysis of data from iCLIP and derived methods, such as eCLIP[16], generally refer to the nucleotide preceding the iCLIP read on the reference genome as the 'crosslink site'. However, in spliceosome iCLIP we additionally expect cDNAs that truncate at the three-way junction formed by intron lariats, where the 5' end of the intron is linked via a 2'-5' phosphodiester bond to the BP (Fig. 2c). Following RNase digestion, such lariat three-way-junction RNAs present two available 3' ends for ligation of adapters, such that cDNAs can truncate at the BP (i.e. position B) or at the start of the intron (i.e. position A). Interestingly, the medium purification condition was optimal to produce cDNAs truncating at positions A and B (Supplementary Fig. 2a), possibly because spliceosomal C complexes containing lariat intermediates are known to be stable under high-salt conditions[17]. Note that peaks A and B are higher in HEK293 compared to HepG2 and K562 cells under medium purification conditions, and likely reflect differences in lariat co-purification. Meanwhile, the number of cDNAs truncating at the positions A and B is dramatically decreased under conditions that inhibit splicing progression and lariat formation: PRPF8 KD *in vivo* (2-fold, Fig. 2a), or absence of ATP *in vitro* (≥18-fold, Fig. 2b). This further confirms that spliceosome iCLIP can monitor spliceosome assembly at distinct stages of the splicing cycle.

Specific RBPs are enriched at each peak of spliceosomal crosslinking

Next, to identify RBPs that crosslink at peaks identified by spliceosome iCLIP, we examined the eCLIP data for 110 RBPs (from 157 eCLIP samples of 68 RBPs in the HepG2, and 89 RBPs in the K562 cell line) provided by the ENCODE consortium[16]. Of note, comparisons between iCLIP and eCLIP are justified due to their use of identical lysis and wash buffers (analogous to medium stringency from the present study), use of truncated cDNAs to identify crosslink sites and similar RNase digestion conditions, and comparable crosslinking profiles for RBPs such as PTBP1 and U2AF2[15]. Accordingly, we analyzed the eCLIP data to identify RBPs with enriched normalized crosslinking at each spliceosomal iCLIP peak. This identified a specific set of RBPs at each peak, with good overlap between RBPs identified in K562 and HepG2 cells (Fig. 3 and Supplementary Data Set 1). As expected, SF3 components SF3B4, SF3A3 and SF3B1 bind to peaks 4 or 5[18], U2AF2 binds the polypyrimidine (polyY) tract (peak 6), and U2AF1 close to the intron-exon junction (peak 7)[19].

Spliceosome iCLIP identifies BPs with canonical sequence and structural features

To determine whether spliceosome iCLIP could experimentally identify human BPs, we used spliceosome iCLIP data produced under medium purification from Cal51 cells. Most cDNA starts in spliceosome iCLIP overlap with a uridine-rich motif (Fig. 4a), in agreement with an increased propensity of protein-RNA crosslinking at uridine-rich sites[14]. In contrast, cDNAs ending at the last nucleotide of introns, which are thus likely derived from intron lariats, have starts overlapping the YUNAY motif matching the consensus BP sequence (Fig. 4b). Further, these cDNAs have higher enrichment of mismatches of adenosines at their first nucleotide (Supplementary Fig. 3a), which is consistent with mismatch, insertion and deletion errors during reverse transcription across the three-way junction of the BP[9]. For comparison, reads that start in regions where BPs are typically located, but which do not align with intron ends, have less enrichment of the BP consensus motif at their starts (Supplementary Fig. 3b,c). To identify a confident set of putative BPs in a transcriptome-wide manner, we therefore used the spliceosome iCLIP cDNAs that aligned with the end of introns (Fig. 4b). These cDNAs started at adenines in 132,287 intronic positions, which we considered as BP candidates. The 41 read-length limited our analysis to the region where most BPs are located, but more distal BPs cannot be identified by this approach. For further study, we selected BPs with the highest number of truncated cDNAs per intron. This identified candidate BPs in 43,637 introns of 9,565 genes.

To examine the BPs identified by spliceosome iCLIP ('iCLIP BPs'), we compared them with the 'computational BPs' recently identified with a sequence-based deep learning predictor, LaBranchoR, which predicted BPs for over 90% of 3'ss[12]. We also compared with 'RNA-seq BPs', including the 138,314 BPs from 43,637 introns that were identified by analysis of lariat-spanning reads from 17,164 RNA-seq datasets[8]. Initially, 65% of iCLIP BPs overlapped with the top-scoring computational BPs (Supplementary Fig. 3d). Interestingly, in cases where iCLIP and computational BPs were located <5 nt apart, they frequently occurred within A-rich sequences (Supplementary Fig. 3e). This mismatch could be of technical nature, as truncation of iCLIP cDNAs may not always be precisely aligned to the BPs in case of A-rich sequences. Alternatively, more than one A might be capable of serving as the BP. When allowing a 1 nt shift for comparison between methods, as has been done previously[12], 70% of iCLIP BPs overlapped with the top-scoring computational BPs, whilst 26% overlapped with the RNA-seq BPs (Fig. 4c, Supplementary Data Set 2). If the computational BPs overlapped either with an iCLIP BP and/or RNA-seq BP, it generally had a strong BP consensus motif (o-BP, Fig. 4d).

To gain insight into the differences between the methods, we focused on BPs that were identified by a single method and located >5 nt away from BPs identified by other methods. Notably, the computational- or iCLIP-specific BPs have a strong enrichment of the consensus YUNAY motif (c-BP, i-BP, Fig. 4e,f,h,i). In contrast, RNA-seq-specific BPs contain a larger proportion of non-canonical BP motifs, which agrees with previous observations[7,9,12] (Fig. 4g,j). To evaluate further, we compared iCLIP BPs with two studies that identified 59,359 BPs by exoribonuclease digestion and targeted RNA-sequencing[9], and 36,078 BPs by lariat-spanning reads refined by U2 snRNP/pre-mRNA base-pairing models[7]. Considering the introns that contained BPs defined both by RNA-seq and iCLIP, we found 57% and 47% overlapping BPs (Supplementary Fig. 3f-i). Again, the iCLIP-specific BPs were more strongly enriched in the consensus YUNAY motif compared to BPs specifically identified by either RNA-seq method (Supplementary Fig. 3j-o). We also examined the local RNA structure around each category of BPs. Overlapping, iCLIP-specific and computational-specific BPs had a decreased pairing probability at the position of the BP, which was not seen for the RNA-seq-specific BPs (Fig. 4k,l). The difference in RNA-seq BPs derives from the presence of non-canonical, non-A branched BPs, which have a generally increased pairing probability (Supplementary Fig. 3p,q). This indicates that the non-A BPs might be structurally less accessible for pairing with U2 snRNP.

Alignment of RBP binding profiles signifies the functionality of BPs

Peaks 4, 5 and position B align to BP position, and therefore we could evaluate how the crosslinking profiles of RBPs binding at these peaks align to the different classes of BPs. First, we examined the crosslinking of SF3B4, which binds in the region of peak 4 as part of the U2 snRNP complex that recognises the BP[1]. Analysis of the overlapping BPs (o-BP) defines the peak of SF3B4 crosslinking at the 25th nt upstream of BPs (Fig. 5 and Supplementary Fig. 4a,b). However, the peak of SF3B4 crosslinking is shifted from this 25th position for the non-overlapping, method-specific BPs; it is generally closer than 25 nt to the BPs located upstream of another BP (up BP), and further than 25 nt away from BPs located downstream of another BP (down BP) (Fig. 5). The shift from the expected position is greatest for RNA-seq-specific BPs (R-BP), and smallest for computationally predicted BPs, as evident by eCLIP data from two cell lines (Fig. 5a,b). Moreover, the same result is seen with U2AF2, where the strongest shift away from expected positions is seen for RNA-seq BPs, and weakest for computational BPs (Supplementary Fig. 4c,d). The cDNA starts from PRPF8 eCLIP are highly enriched at position B, corresponding to the lariat-derived cDNAs that truncate at BPs (Fig. 3). Interestingly, the PRPF8 cDNA starts had the strongest peak at the overlapping BPs, but also peaked at all the remaining classes of BPs (Supplementary Fig. 4e,f). This indicates that all classes of BPs contribute to lariat formation, and that the non-overlapping BPs most likely act as alternative BPs within the introns.

Effects of BP position on spliceosomal assembly

To assess how BP positioning determines spliceosome assembly, we evaluated binding profiles of the RBPs that are enriched at peaks 4-7 and at positions A and B (Fig. 3). We divided BPs based on their distance from 3'ss, and normalized RBP binding profiles within each subclass of BP. This showed that crosslinking of U2AF1 and U2AF2 aligns to the region between the BPs and 3'ss, which is covered by the polyY tract (Supplementary Fig. 5 and 6). Whilst SF3B4 is the primary RBP crosslinking at peak 4, and SF3A3 at peak 5, binding of SMNDC1, SF3B1, EFTUD2, BUD13, GPKOW and XRN2 to peaks 4 and 5 was also evident (Supplementary Fig. 5, 6 and Fig. 3). PRPF8, RBM22 and SUPV3L1 have their cDNA starts truncating at positions A and B (Supplementary Fig. 5 and 6), corresponding to the three-way junction formed by intron lariats (Fig. 2c). This is in agreement with the association of PRPF8 and RBM22 with intron lariats as part of the human catalytic step I spliceosome[1]. The positions of SF3B4 and SF3A3 crosslinking peaks also agree with CryoEM studies of the human spliceosome that show closer pre-mRNA binding of SF3A3 (also referred to as SF3a60) to the BP compared to SF3B4 (also referred to as SF3b49)[20].

In order to quantify how BP positioning affects the intensity of RBP binding, we divided BPs into 10 equally sized groups based on the distance from 3'ss. We then normalized the relative binding intensity of each RBP at each position on the RNA maps across the ten groups, and revealed strong relationships between BP position and binding intensity of certain RBPs (Fig. 6a, Supplementary Fig. 7a). For example, if a BP is located distally from the 3'ss, then U2AF components bind stronger to peaks 6 and 7. In contrast, if a BP is located proximally to the 3'ss, then EFTUD2, SF3 components and several other RBPs bind stronger to the peaks 4 or 5 (Fig. 6b). Notably, increased BP distance causes increased binding of BUD13 and GPKOW at peaks 6 or 7 and decreased binding at peaks 4 and 5. The more efficient recruitment of U2AF and associated factors to peaks 6 and 7 could be explained by the long polyY-tracts at distal BPs (Supplementary Fig. 5), while their decreased binding at proximal BPs appears to be compensated by increased binding of SF3 and other U2 snRNP-associated factors at peaks 4 and 5.

In contrast to effects on individual splicing factors, we did not observe any effect of BP distance on the relative intensity of spliceosome iCLIP crosslinking in peaks 4 and 5 compared to 6 and 7 (Fig. 6c). This indicates that the effects may be masked during later stages of spliceosome assembly. To ask if this is the case, we turned to PRPF8, a protein that is essential for later stages of spliceosomal assembly, a role it plays together with EFTUD2 and BRR2 as part of U5 snRNP[1]. PRPF8 KD leads to decreased spliceosomal binding at peaks 4 and 5, and this effect is stronger at distal compared to proximal BPs (Fig. 6c). In conclusion, our results reveal differences in the binding profiles of splicing factors in relation to BP distance, but these differences are neutralized upon full spliceosome assembly in a manner that requires the presence of PRPF8.

Effects of BP strength on spliceosomal assembly

To examine how BP strength affects spliceosomal assembly we focused on BPs that have been identified both by spliceosome iCLIP and computational modelling, and which are located at 23-28 nt upstream of the 3'ss. Of note, this is the most common position of BPs (Supplementary Data Set 3). As an estimate of BP strength we used the BP score, which was determined with a deep-learning model[12]. This showed strong correlation between BP strength and RBP binding intensities, such that most RBPs have increased crosslinking at peaks 4 and 5 at BPs with very high scores, and, conversely, increased crosslinking at peaks 6 and 7 at BPs with very low scores (Fig. 7a,b, Supplementary Fig. 7b). Since SF3 components primarily bind at peaks 4 and 5, and U2AF components at peaks 6 and 7, an over 4-fold change is seen in the ratio of crosslinking when comparing the extreme deciles of BP strength (Supplementary Fig. 7c). We did not observe any correlation between the polyY tract coverage and BP score (Supplementary Fig. 7d), which indicates that BP strength directly affects the RBP binding profiles.

Similar to the effects on individual splicing factors, the relative intensity of spliceosome iCLIP crosslinking in peaks 4 and 5 was increased with increasing BP strength (Fig. 7c, compare blue lines on the left and right graphs). PRPF8 KD decreased spliceosomal binding at peaks 4 and 5 of both classes of BPs, and this led to stronger crosslinking at peaks 6 and 7 relative to peaks 4 and 5 at weak BPs, even though the peaks 4 and 5 are usually stronger. The signal at position B of weak BPs is almost completely lost upon PRPF8 KD, which likely reflects the absence of intron lariats due to perturbed splicing of introns with weak BPs (Fig. 7c). In conclusion, our results suggest that the assembly efficiency of spliceosomal factors at peaks 4 and 5 closely correlates with BP strength, which indicates that recognition of weak BPs might be more sensitive to perturbed spliceosome function.

Discussion

Here we established spliceosome iCLIP to study the interactions of endogenous snRNPs and accessory splicing factors on pre-mRNAs. We identified peaks of spliceosomal protein-pre-mRNA interactions, which precisely overlap with crosslinking profiles of 15 splicing factors. Interestingly, the contacts of RBPs in peaks 4 and 5 don't overlap with any sequence motif, and thus the constrained conformation of the larger spliceosomal complex appears to act as a molecular ruler that positions each associated RBP on pre-mRNAs at a specific distance from BPs. Moreover, the presence of lariat-derived reads in spliceosome iCLIP identified >40,000 BPs that have canonical sequence and structural features. Due to the precise alignment of splicing factors relative to the positions of BPs, we could use their binding profiles to show that the assembly of U2 snRNP is primarily coordinated by the computationally predicted BPs, whilst alternative BPs, identified only by iCLIP or RNA-seq, are more rarely used. Finally, we reveal the major effect of the position and strength of BPs on spliceosomal assembly, which can explain why distally located or weak BPs are particularly sensitive to perturbed spliceosome function upon PRPF8 KD. These findings demonstrate the broad utility of spliceosome iCLIP for simultaneous and transcriptome-wide analysis of the assembly of diverse spliceosomal components.

The value of spliceosome iCLIP for identifying BPs

Both RNA-seq and iCLIP identify BPs by analyzing cDNAs derived from intron lariats. Thus, the efficiency of these methods depends on the abundance of intron lariats, which depends on the kinetics of lariat debranching. Several studies demonstrated that lariats formed at non-canonical BPs are less efficiently debranched[21-23], and therefore these non-canonical BPs are expected to be more efficiently detected. This is especially true for RNA-seq-based methods, because they monitor steady state RNA levels. In contrast, iCLIP only captures lariats in complex with spliceosomes, thus minimizing bias for lariats that are stable after their release from the spliceosome. This could explain why the BPs identified by iCLIP contain a stronger consensus sequence than BPs identified from lariat-spanning reads in RNA-seq. The further value of spliceosome iCLIP is that, in addition to experiments under the medium condition that permit BP identification through lariat-derived cDNAs, experiments under the mild condition identify the SF3 complex and other U2 snRNP-associated RBPs that crosslink at peaks 4 and 5. These can crucially be used to independently validate the functional role of BPs in the assembly of U2 snRNP. Thus, use of spliceosome iCLIP under both conditions, combined with computational modelling of BPs[12], is well suited to studying the functionality of BPs.

The role of BP position and strength in spliceosomal assembly

We show that BP position and the computationally defined strength of BPs correlate with the relative binding of splicing factors around BPs. This is exemplified by strong binding of SF3 components at strong BPs, or BPs located close to 3'ss, whilst U2AF components bind stronger to weak BPs, or BPs located further from 3'ss (Fig. 7d). In the cases of SF3B1, BUD13 and GPKOW, we observed enriched binding at peaks 4 and 5 as well as 6 and 7, with reciprocal changes between the two peak regions dependent on BP features (Fig. 6 and 7). These RBPs are not known to bind at peaks 6 or 7, and it is plausible that the signal at some peaks represents binding of U2AF or other spliceosomal factors that are co-purified during eCLIP. It is presently not possible to fully distinguish between direct and indirect binding from eCLIP data, because purified protein-RNA complexes have not been visualized after their separation on SDS-PAGE gels in eCLIP[13]. Nevertheless, it is clear that BP characteristics determine the balance between binding of SF3 and associated factors at peaks 4 and 5 and of U2AF and associated factors at peaks 6 and 7. This suggests further study of RBP binding profiles around BPs could unravel a BP ‘code’ that facilitates specific stages of BP recognition and function.

In conclusion, spliceosome iCLIP monitors concerted pre-mRNA binding of many types of spliceosomal complexes with nucleotide resolution, allowing their simultaneous study due to the distinct position-dependent binding pattern of components acting at multiple stages of the splicing cycle. The method can now be used to study the endogenous spliceosome and BPs across tissues, species and stages of development without need for the protein tagging used in yeast[3,4]. Further, several spliceosomal components, including U2AF1, SF3B1 and PRPF8, are targets for mutations in myeloid neoplasms, retinitis pigmentosa and other diseases[24]. Spliceosome iCLIP could now be used to monitor global impacts of these mutations on spliceosome assembly in human cells. More generally, our study demonstrates the value of iCLIP for monitoring the position-dependent assembly and dynamics of multi-protein complexes on endogenous transcripts.

Acknowledgements

We thank M. Llorian for help with the *in vitro* splicing reactions, K. Zarnack and G. Rot for help with the data analyses, and L. Strittmatter and members of Ule lab for helpful discussions and comments on the manuscript. This work was supported primarily by the European Research Council (206726-CLIP and 617837-Translate) and the Slovenian Research Agency (P2-0209, Z7-3665, J7-5460). C.R.S. was supported by an Edmond Lily Safra fellowship, and a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant number 215454/Z/19/Z). A.S.E. is supported by the Biotechnology and Biological Sciences Research Council (BB/M009513/1). A.M.C. is supported by a Wellcome Trust PhD Training Fellowship for Clinicians (110292/Z/15/Z). D.P. and V.O.W. were supported by Medical Research Council programme grants MC_UU_12022/1 and MC_UU_12022/8 to A.R.V.. L.P. was supported by NIH-NINDS (R01 NS102451). The Francis Crick Institute receives its core funding from Cancer Research UK (FC001002), the UK Medical Research Council (FC001002), and the Wellcome Trust (FC001002).

Author contributions

M.B., C.R.S. and J.U. conceived the project, designed the experiments and wrote the manuscript, with assistance of all co-authors. M.B., C.R.S., Z.W., R.F. and A.S.E. performed experiments, with assistance from J.U., J.K. and C.W.S.. N.H. performed most computational analyses, with assistance from C.R.S., T.C., R.F., A.M.C. and N.M.L.. V.O.W., D.P. and A.R.V. provided crosslinked pellets from wild-type and PRPF8-depleted Cal51 cells. L.S. and L.P. developed and characterized the monoclonal antibody 18F6.

Competing interests

The authors declare no competing interests.

References

1. Fica, S.M. & Nagai, K. Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nat Struct Mol Biol* **24**, 791-799 (2017).
2. Wahl, M.C., Will, C.L. & Lührmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701-18 (2009).
3. Chen, W. et al. Transcriptome-wide Interrogation of the Functional Intronome by Spliceosome Profiling. *Cell* **173**, 1031-1044 e13 (2018).
4. Burke, J.E. et al. Spliceosome Profiling Visualizes Operations of a Dynamic RNP at Nucleotide Resolution. *Cell* **173**, 1014-1030 e17 (2018).
5. Wickramasinghe, V.O. et al. Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5' splice site strength. *Genome Biol* **16**, 201 (2015).
6. König, J. et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**, 909-15 (2010).
7. Taggart, A.J. et al. Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res* **27**, 639-649 (2017).
8. Pineda, J.M.B. & Bradley, R.K. Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev* **32**, 577-591 (2018).
9. Mercer, T.R. et al. Genome-wide discovery of human splicing branchpoints. *Genome Res* **25**, 290-303 (2015).
10. Carissimi, C., Saieva, L., Gabanella, F. & Pellizzoni, L. Gemin8 is required for the architecture and function of the survival motor neuron complex. *J Biol Chem* **281**, 37009-16 (2006).
11. Huppertz, I. et al. iCLIP: protein-RNA interactions at nucleotide resolution. *Methods* **65**, 274-87 (2014).
12. Paggi, J.M. & Bejerano, G. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA* **24**, 1647-1658 (2018).
13. Lee, F.C.Y. & Ule, J. Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Mol Cell* **69**, 354-369 (2018).
14. Sugimoto, Y. et al. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome biology* **13**, R67 (2012).
15. Haberman, N. et al. Insights into the design and interpretation of iCLIP experiments. *Genome Biol* **18**, 7 (2017).
16. Van Nostrand, E.L. et al. A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv* (2017).
17. Bessonov, S., Anokhina, M., Will, C.L., Urlaub, H. & Luhrmann, R. Isolation of an active step I spliceosome and composition of its RNP core. *Nature* **452**, 846-50 (2008).
18. Gozani, O., Feld, R. & Reed, R. Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A. *Genes Dev* **10**, 233-43 (1996).
19. Zarnack, K. et al. Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell* **152**, 453-66 (2013).

20. Zhang, X. et al. Structure of the human activated spliceosome in three conformational states. *Cell Res* **28**, 307-322 (2018).
21. Jacquier, A. & Rosbash, M. RNA splicing and intron turnover are greatly diminished by a mutant yeast branch point. *Proc Natl Acad Sci U S A* **83**, 5835-9 (1986).
22. Hesselberth, J.R. Lives that introns lead after splicing. *Wiley Interdiscip Rev RNA* **4**, 677-91 (2013).
23. Talhouarne, G.J.S. & Gall, J.G. Lariat intronic RNAs in the cytoplasm of vertebrate cells. *Proc Natl Acad Sci U S A* **115**, E7970-E7977 (2018).
24. Scotti, M.M. & Swanson, M.S. RNA mis-splicing in disease. *Nat Rev Genet* **17**, 19-32 (2016).
25. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011).

Figure legends

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

(a) Schematic representation of the spliceosome iCLIP method performed under conditions of varying purification stringency.

(b) Autoradiogram of crosslinked RNPs immunopurified from HeLa cells under medium conditions by a SmB/B' antibody following digestion with high (++) or low (+) amounts of RNase I. The dotted line depicts the region typically excised from the nitrocellulose membrane for spliceosome iCLIP. As control, the antibody (Ab) was omitted during immunopurification.

(c) Genomic distribution of spliceosome iCLIP cDNAs produced under stringent, medium and mild conditions from HEK293 cells. Data was mapped first to snRNAs, allowing multiple mapping reads, and then to the genome, allowing only uniquely mapped reads. Proportions of cDNAs mapping to snRNAs, introns, coding sequence of mRNAs (CDS), untranslated regions of mRNAs (UTR) and long non-coding RNAs (lncRNAs) are shown (but not the intergenic reads and other types of RNAs). Data are shown as mean±s.e.m from three independent experiments for the medium and mild purification condition and two independent experiments for the stringent purification condition. Source data for panel c are available online.

Fig. 2 | Analysis of spliceosomal interactions with pre-mRNAs *in vitro* and *in vivo*.

(a) Metagene plots of spliceosome iCLIP from Cal51 cells. Plots are depicted as RNA maps of summarized crosslinking at all exon-intron and intron-exon boundaries, and around BPs to identify major binding peaks, and to monitor changes between control and PRPF8 knockdown (KD) cells. Crosslinking is regionally normalized to its average crosslinking across the -100..50 nt region relative to splice sites or BPs depending on the RNA map in order to focus the comparison on the relative positions of peaks.

(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 crosslinking peaks 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. 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'ss and truncate at the first nucleotide of the intron (position A).

Fig. 3 | Identification of RBPs overlapping with spliceosomal peaks at BPs and 3'ss.

Enrichment of eCLIP crosslinking within each of the spliceosome iCLIP peaks, which are defined by the positions marked in the figure. We first regionally normalized the crosslinking of each RBP to its average crosslinking over -100..50 nt region relative to 3'ss, which generates the RNA maps as shown in Supplementary Fig. 5 and 6. We then ranked the RBPs according to the average normalized crosslinking across the nucleotides within each peak. We analyzed peaks 4-7 and positions A and B, as marked on the top of each plot. The top-ranking RBPs in each peak are shown on the left plot, and the full distribution of RBP enrichments is shown on the right plot.

Fig. 4 | Comparison of BPs identified by spliceosome iCLIP, RNA-seq lariat reads or computational prediction.

(a) Weblogo around the nucleotide preceding all spliceosome iCLIP reads.

(b) Weblogo around the nucleotide preceding only those spliceosome iCLIP reads that align with ends of introns.

(c) Introns that contain at least one BP identified either by published RNA-seq[8] or by spliceosome iCLIP are used to examine the overlap between the top BPs identified by RNA-seq (i.e., the BP with most lariat-spanning reads in each intron), iCLIP (BP with most cDNA starts) or computational predictions (highest scoring BP)[12]. BPs that are 0 or 1 nt apart are considered as overlapping. At the right, BP categories that are used for all subsequent analyses are defined, along with their acronyms. If a BP defined by one method is >5 nt upstream of a BP defined by another method, then 'up' is added to its acronym, and if it is >5 nt downstream, 'down' is added.

(d) Weblogo of o-BP category of BPs.

(e) Weblogo of C-BPup category of BPs.

(f) Weblogo of i-BPup category of BPs.

(g) Weblogo of R-BPup category of BPs.

(h) Weblogo of C-BPdown category of BPs.

(i) Weblogo of i-BPdown category of BPs.

(j) Weblogo of R-BPdown category of BPs.

(k, l) The 100 nt RNA region centered on the BP was used to calculate pairing probability with the RNAfold program using default parameters[25], and the average pairing probability of each nucleotide around BPs is shown for the 40 nt region around method-specific BPs located upstream (k) or downstream (l).

Fig. 5 | Spliceosome assembly at BPs identified by spliceosome iCLIP, RNA-seq lariat reads or computational prediction.

Violin plots depicting the positioning of SF3B4 cDNA starts relative to the indicated BP categories. SF3B4 eCLIP data were from K562 (a) and HepG2 (b) cells. Box-plot elements are defined by center line, median; box limits, upper and lower quartiles; and whiskers, 1.5× interquartile range. Each data point corresponds to an eCLIP crosslink event, and the total number of eCLIP crosslinks that map in the area analysed around each set of BPs (sample size) is shown under the plot.

Fig. 5 | Spliceosome assembly at BPs identified by spliceosome iCLIP, RNA-seq lariat reads or computational prediction.

Violin plots depicting the positioning of SF3B4 cDNA starts relative to the indicated BP categories. SF3B4 eCLIP data were from K562 (a) and HepG2 (b) cells. Box-plot elements are defined by center line, median; box limits, upper and lower quartiles; and whiskers, 1.5× interquartile range.

Fig. 6 | BP position defines the binding patterns of splicing factors at 3'ss.

(a) Heatmaps depicting the normalized crosslinking of RBPs in peak regions around 10 groups of BPs that were categorized according to the distance of the BP from 3'ss. Crosslinks were derived as cDNA starts from eCLIP of HepG2 cells.

(b) RNA maps showing normalized crosslinking profiles of selected RBPs relative to BPs and 3'ss for the two deciles of BPs that are located most proximal (interrupted light lines) or most distal (solid dark lines) from 3'ss.

(c) RNA maps showing crosslinking profile of spliceosome iCLIP from control and PRPF8 KD Cal51 cells in the same format as panel b.

Fig. 7 | BP strength correlates with the binding of splicing factors.

(a) Heatmaps depicting the normalized crosslinking of RBPs in peak regions around 10 groups of BPs that were categorized according to the computational scores that define BP strength. Crosslinks were derived as cDNA starts from eCLIP of HepG2 cells.

(b) RNA maps showing normalized crosslinking profiles of selected RBPs relative to 3'ss for the two deciles of BPs that are lowest scoring (interrupted light lines) or highest scoring (solid dark lines).

(c) RNA maps showing crosslinking profile of spliceosome iCLIP from control and PRPF8 KD Cal51 cells in the same format as panel b.

(d) Schematic representation of the effects that BP position and score have on the assembly of SF3 and U2AF complexes around BPs.

Online Methods

Cell culture

Flp-In HEK293 T-REx cells were from ThermoFisher (R78007), K562, HepG2 and standard HEK293 cells were obtained from the Francis Crick Cell Services Science Technology Platform, and Cal51 breast adenocarcinoma cells were obtained from DSMZ (reference 14563). All cell lines tested negative for Mycoplasma contamination. HEK293 and HepG2 were cultured in DMEM with 10% FBS (ThermoFisher) and 1× penicillin-streptomycin (ThermoFisher). K562 cells were cultured in RPMI 1640 (IMDM, ATCC) with 10% FBS and 1× penicillin-streptomycin. Cal51 cells were cultured in DMEM (ThermoFisher) with 10% fetal calf serum (FCS, ThermoFisher) and 1× penicillin-streptomycin (ThermoFisher).

To generate a plasmid encoding 3×Flag epitope-tagged SmB, the SmB cDNA was amplified using Phusion High-Fidelity DNA polymerase (NEB) with primers carrying the KpnI and NotI restriction enzymes sites and cloned using Rapid DNA Ligation Kit (Thermo Fisher Scientific) into a pcDNA5/FRT/TO vector modified to encode 3×Flag peptide upstream of the multiple cloning site. To produce stable cell lines expressing this construct, the pcDNA5/FRT/TO plasmid with 3×Flag epitope-tagged SmB was co-transfected with pOG44 plasmid into Flp-In HEK293 T-REx cells (ThermoFisher, R78007). Cells stably expressing these proteins were selected by culturing in Dulbecco's Modified Eagle Medium (DMEM, Thermofisher) containing 10% fetal bovine serum (FBS), 3 µg/ml Blasticidine S HCl, 200 µg/ml Hygromycine (InvivoGen). Flp-In 293 T-REx cells (Life Technologies) were cultured in DMEM with 10% FBS, 3 µg/ml Blasticidin S HCl (Life Technologies), 50 µg/ml Zeocin (Life Technologies). Doxycycline was added to media 24 hours prior to sample preparation in order to induce construct expression.

Cal51 breast adenocarcinoma cells were prepared as described previously[5]. For siRNA-mediated depletion of PRPF8, Cal51 cells were transfected using DharmaFECT1 (Dharmafect) with 25 nM siRNA targeting human *PRPF8*. Transfected cells were harvested 54 hrs later, exposed to UV-C light and used for iCLIP as described below. For collection of samples from different stages of the cell cycle, Cal51 cells were synchronized in G1/S by standard double thymidine block. Briefly, cells were treated with 1.5 mM thymidine for 8 hrs, washed and released for 8 hrs, then treated again with thymidine for a further 8 hrs. Cells were also collected 3 hrs (S-phase) and 7 hrs (G2) after release from the thymidine block.

Antibody production

For production of the anti-SmB/B' monoclonal antibody 18F6, Balb/c females were primed with Immuneasy adjuvant (Qiagen) and 25 mg of 6×His-SmB purified recombinant proteins. Following two boosts at two-week intervals, SP2 myeloma cells were fused with mouse splenocytes and hybridoma supernatants were analyzed onto antigen-coated aminosilane modified slides using a LS400 Scanner (Tecan) and the GenePix Pro 4.1 software as described previously[10]. Hybridoma cells were subcloned by limiting dilution and further screened by ELISA, Western blot and immunofluorescence analysis of HeLa cells.

In vitro splicing

For *in vitro* splicing reactions, a *C6orf10* minigene construct containing exon 8 and 9 and 150 nt of the intron around both splice sites was produced (Fig. 2b). The minigene plasmid was linearized and transcribed *in vitro* using T7 polymerase with ^{32}P-UTP. The transcribed RNA was then subjected to *in vitro* splicing reactions using HeLa nuclear extract. HeLa nuclear extract was depleted of endogenous ATP by pre-incubation and, for each reaction, 10 ng of RNA was incubated with 60% HeLa nuclear extract at 30°C with or without additional 0.5 mM ATP for 1 h in a 20 µl reaction. Afterwards, the reaction mixture was UV-crosslinked at 100 mJ/cm^{2} and stored at -80°C until further use. To visualize the splicing reaction products, proteinase K was added to the reaction mixture for 30 min at 37°C. The resulting RNA was phenol-extracted, precipitated and subjected to gel electrophoresis on a 5% polyacrylamide-urea gel.

Spliceosome iCLIP protocol

For each experiment, three biological replicate samples of cDNA libraries were prepared (Supplementary Tables 2 and 3). The iCLIP method was done as previously described[11], with the following modifications. Crosslinked cells or tissue were dissociated in the lysis buffer according to the stringency conditions (stringent, medium, mild; Supplementary Table 1) followed by sonication, low RNase I (AM2295, 100 U/µl, ThermoFisher) digestion and centrifugation. RNase at low concentration ensured that cDNAs are of optimal size for comprehensive crosslink determination[15]. For denaturing, high-stringency experiment[11], M2 anti-Flag antibody (Sigma) was used against the 3×Flag-SmB protein that had been stably integrated into HEK-293 FlpIn cells (Supplementary Fig. 1c). 6M Urea buffer was first used to lyse cell pellets, before being diluted down 1:9 with a Tween-20-containing IP buffer to allow for immunopurification without denaturing of the M2 anti-Flag antibody, and then proceeded as described previously[15].

Standard iCLIP protocol[11] was used for Cal51 cells under mild and medium stringency conditions, and for the *in vitro* splicing reactions under mild conditions, whilst an updated protocol was used for HEK293, HepG2 and K562 cells[26]. For SmB/B' immunopurification anti-SmB/B' antibodies 12F5 (sc-130670, Santa Cruz Biotechnology for Cal51 cells, and S0698, Sigma-Aldrich for HEK293, HepG2 and K562 cells) or 18F6 (as hybridoma supernatant, generated as described previously[10]) were used, which are different clones from the same immunization. These antibodies behave identically under immunopurification conditions (Supplementary Fig. 1d). For spliceosome iCLIP from *in vitro* splicing reactions (Supplementary Fig. 2c,d), lysates were incubated with 50 µl monoclonal anti-SmB/B' antibody 18F6, and for immunoprecipitations from cell lysates, 12F5 anti-SmB/B' antibody was used. The antibody was bound to 100 µl protein G Dynabeads (ThermoFisher) under rotation at 4°C followed by washing. As described previously, following immunopurification, RNA 3' end dephosphorylation, ligation of the adapter 5'-rAppAGATCGGAAGAGCGGTTCAG/ddC/-3' to the 3' end and 5' end radiolabeling, protein-RNA complexes were size-separated by SDS-PAGE and transferred onto nitrocellulose membrane. The regions corresponding to 28-180 kDa were excised from the membrane in order to isolate the bound RNA by proteinase K treatment. RNAs were reverse-transcribed in all experiments using SuperScript III or IV reverse transcriptase (ThermoFisher) and custom indexed primers (Supplementary Table 2). Resulting cDNAs were subjected to electrophoresis on a 6% TBE-urea gel (ThermoFisher) for size selection. Purified cDNAs were circularized, linearized and amplified for high-throughput sequencing.

Identification of protein crosslink sites around splice sites, in particular at the peaks 4 and 5, was most efficient under the mild purification condition (Supplementary Fig. 2a). This condition was therefore used for analysis of spliceosomal assembly upon PRPF8 knockdown in Cal51 cells (Fig. 2a), and in the *in vitro* splicing reactions in HeLa nuclear extract (Fig. 2b). For the identification of BPs, we additionally used the medium condition, since it increases the frequency of cDNAs truncating at peak B (Supplementary Fig. 2a). For this purpose, spliceosome iCLIP was performed under medium purification conditions from Cal51 cells synchronized in G1, S and G2 phase. To maximize cDNA coverage, data from all synchronized cells was merged with the control Cal51 cells under mild condition for BP identification.

Mapping of Sm iCLIP reads

We mapped iCLIP data to the GRCh38 primary assembly and GENCODE v27 gene annotations using STAR (v.2.2.1). Experimental and random barcode sequences of iCLIP sequenced reads were removed prior to mapping (Supplementary Table 2). Following mapping, we used random barcodes to quantify the number of unique cDNAs at each genomic position by collapsing cDNAs with the same random barcode that mapped to the same starting position to a single cDNA. For analysis of crosslinking to snRNAs, we first mapped to a transcriptome of all annotated snRNA sequences in GENCODE v27 using Bowtie2 (v2.3.4.3), and kept the primary alignment. Unmapped reads were then mapped with STAR as previously described and intersected with GENCODE v27 for subtype analysis, with reads from Bowtie2 being added to the total snRNA count. For spliceosome iCLIP with the *C6orf10 in vitro* splicing substrate, sequence reads were first mapped to the unspliced substrate and the remaining reads were mapped to the spliced substrate allowing no mismatches. The nucleotide preceding the iCLIP cDNAs was used to define the crosslink sites in all analyses.

Mapping of eCLIP reads

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

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

Normalization of crosslink positions for their visualization in the form of RNA maps

RNA maps and heat maps were produced by summarizing the cDNA counts at each nucleotide using the previously developed RNA maps pipeline[15,27] relative to exon-intron and intron-exon boundaries and BPs on pre-mRNAs. The definition of intronic start and end positions was based on Ensembl version 75. Only introns longer than 300 nt were used to draw RNA maps in order to avoid detection of any RBPs that recognize 5'ss of introns.

In cases where we wished to compare the relative positions of crosslinking peaks between RBPs, we regionally normalized the summarized crosslinking of each RBP relative to the average crosslinking of the same RBP across the region 100 nt upstream and 50 nt downstream of the evaluated splice sites or BPs. Normalized values were then used to visualize the crosslinking in the form of RNA maps (Fig. 2, Supplementary Fig. 5 and 6). The same normalization was then used to plot heat maps, by plotting mean values of normalized RNA maps for each peak in the following regions; peak 4: -29..-23 nt and peak 5: -21..-17 nt relative to BP, peak 6: -11..-5 nt and peak 7: -3..-1 nt relative to 3'ss. Every RBP was then normalized by the mean across all the peaks to visualize crosslinking enrichment between the groups on the same scale across all RBPs (Fig. 6 and 7, Supplementary Fig. 7).

To assess the role of BP characteristics on spliceosomal RBP assembly (Fig. 4, 6 and 7), we only examined the introns containing the 31,167 BPs that were identified both computationally and by iCLIP, which are likely the most reliable. We divided BPs into 10 categories based on BP position or score, and then normalized the summarized crosslinking of each RBP in each of the 10 BP categories relative to the average crosslinking of the same RBP across the region 100 nt upstream and 50 nt downstream of all the 31,167 evaluated BPs.

For visualization of spliceosome iCLIP crosslinks along the *C6orf10 in vitro* splicing substrate and product (Fig. 2b and Supplementary Fig. 2e) we first summed the cDNA starts at each nt position and then normalized the counts by the average number of cDNA starts in the intronic region 101..150 relative to the 5'ss of the unspliced substrate. For the unspliced substrate normalized cDNA counts were logarithmized ($\log_2$) and data with $\log_2$(normalized number of cDNA starts)≥1 were plotted. For the spliced product normalized cDNA counts were plotted.

Identification and comparison of BPs

It has been shown that the spliceosomal C complexes harbor a salt-resistant RNP core containing U2, U5 and U6 snRNAs as well as the splicing intermediates including lariats that withstand treatment with 1M NaCl, whereas the spliceosomal B complexes were more likely dissociated under high-salt conditions[17]. This could explain why the medium purification condition is more suited than the mild condition to enrich for lariat cDNAs truncating at position B (Supplementary Fig. 2a). It is conceivable that the medium spliceosome iCLIP condition most strongly enriches spliceosomal C complexes, which are most effective for lariat detection. In contrast, the mild condition is expected to enrich additional B complexes that contain large amounts of SF3 components and have low proportion of lariats, in agreement with the strong enrichment of peaks 4 and 5 (Supplementary Fig. 2a). To identify the maximal diversity of BPs, we therefore pooled spliceosome iCLIP data produced under mild and medium purification conditions from Cal51 cells.

To identify BPs we used the spliceosome iCLIP reads that ended precisely at the ends of introns (we considered only introns that end in AG dinucleotide) after removal of the 3' adapter. We noticed that these reads had an 3.5× increased frequency of mismatches on the A as the first nucleotide compared to remaining iCLIP reads (Supplementary Fig. 3a), indicating that these mismatches may have resulted from truncation at the three-way-junction formed at the BP (Fig. 2c). We therefore trimmed the first nucleotide from the read if it contained a mismatch at the first position that corresponded to a genomic adenosine. We then used spliceosome iCLIP from Cal51 cells to identify all reads that ended precisely at the ends of introns and defined the position where these reads started and assessed the random barcode nucleotides that are present at the beginning of each iCLIP read to count the number of unique cDNAs at each position. The nucleotide preceding the read start corresponds to the position where cDNAs truncated during the reverse transcription, and we selected the genomic A that had the highest number of truncated cDNAs as the candidate BP. If two positions with equal number of cDNAs were found, we selected the one closer to the 3'ss. Together, this identified 43,637 BPs.

We also attempted to use truncated cDNAs from PRPF8 eCLIP for discovery of BPs, but found that the number of cDNAs overlapping with intron ends was much smaller than in spliceosome iCLIP, and was insufficient for BP discovery. This is most likely because of the high amount of non-specific background signal in PRPF8 eCLIP, which leads to a lower proportion of cDNAs that align to the BPs.

Bedtools Intersect command using option –u was used to compare BP coordinates from spliceosome iCLIP to the BPs identified in previous studies. We restricted this comparison to introns where BPs were detected by all three datasets (iCLIP, RNA-seq and computational prediction).

To define a single ‘computational BP’ per intron, the BP positions computationally predicted for each intron in hg19 were obtained from http://bejerano.stanford.edu/labranchor/, and the top scoring BP in each intron was used. To define a single ‘RNA-seq BP’ per intron, we used the BP with most lariat-spanning reads in each intron.

Analysis of pairing probability

Computational predictions of the secondary structure were performed by RNAfold function from Vienna Package (https://www.tbi.univie.ac.at/RNA/) with default parameters[25]. The RNAfold results are provided in a customized format, where brackets are representing the double-stranded region on the RNA and dots are used for unpaired nucleotides. We measured the density of pairing probability by summing the paired positions into a single vector.

Identification of RBPs overlapping with spliceosomal peaks

For RBP enrichment in Fig. 3, we used the eCLIP data from the ENCODE consortium[16], together with available iCLIP experiments from our lab (which are all listed in Supplementary Data Set 4), to see if any of the proteins are enriched in the region of spliceosomal peaks. In total, this included 157 eCLIP samples of 68 RBPs in the HepG2 cell line, and 89 RBPs in the K562 cell line, and iCLIP samples of 18 RBPs from different cell lines (Supplementary Data Set 4). Next, we intersected cDNA starts from each sample to the -100 to +50 nt region relative to the 3'ss and used it as control for each of the following peaks: Peak 4 (-23 nt..-29 nt relative to BP), Peak 5 (-21 nt..-17 nt relative to BP), Peak B (-1 nt..1 nt relative to BP), Peak A (-1 nt..1 nt relative to 5'ss), Peak 6 (-11 nt..-10 nt relative to 3'ss), Peak 7 (-3 nt..-2 nt relative to 3'ss). The positions of these peaks were determined based on crosslink enrichments in spliceosome iCLIP.

Statistics

All statistical analyses were performed in the R software environment (version 3.1.3 and 3.3.2, https://www.r-project.org).

Reporting Summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability

The code to identify BPs from spliceosome iCLIP reads is publicly available at the GitHub repository (https://github.com/nebo56/branch-point-detection-2).

Data availability

The spliceosome iCLIP data generated and analyzed during the current study are available on EBI ArrayExpress under the accession number E-MTAB-8182, and are also available in raw and processed format on https://imaps.genialis.com/iclip. Additional datasets used in this study are listed in Supplementary Data Set 4. Source Data for Fig. 1c are available online. Other data are available upon request.

Methods-only references

26. Blazquez, L. et al. Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing. *Mol Cell* **72**, 496-509 e9 (2018).

27. Chakrabarti, A., Haberman, N., Praznik, A., Luscombe, N.M. & Ule, J. Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies. *Annual Review of Biomedical Data Science* **Vol. 1**(2018).

Figure 1

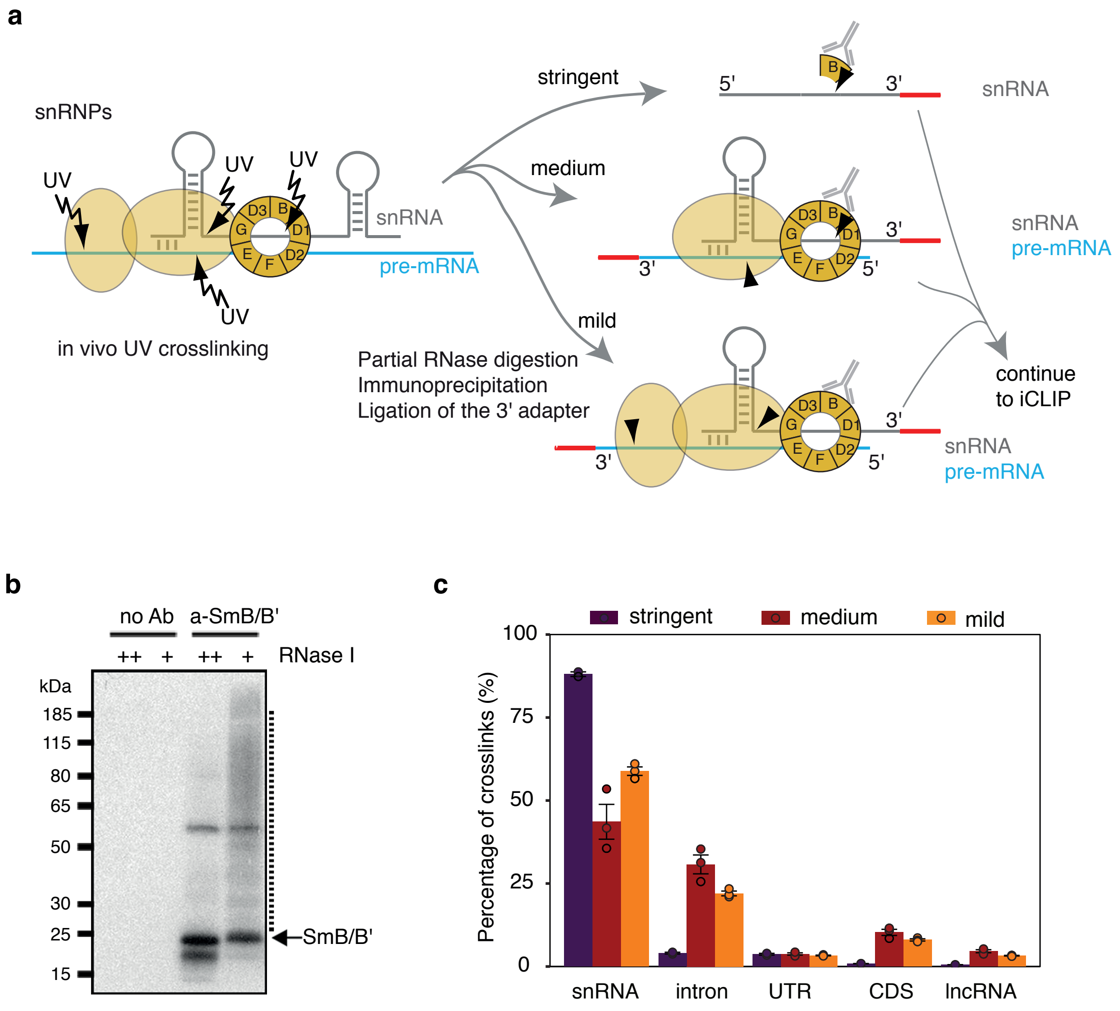
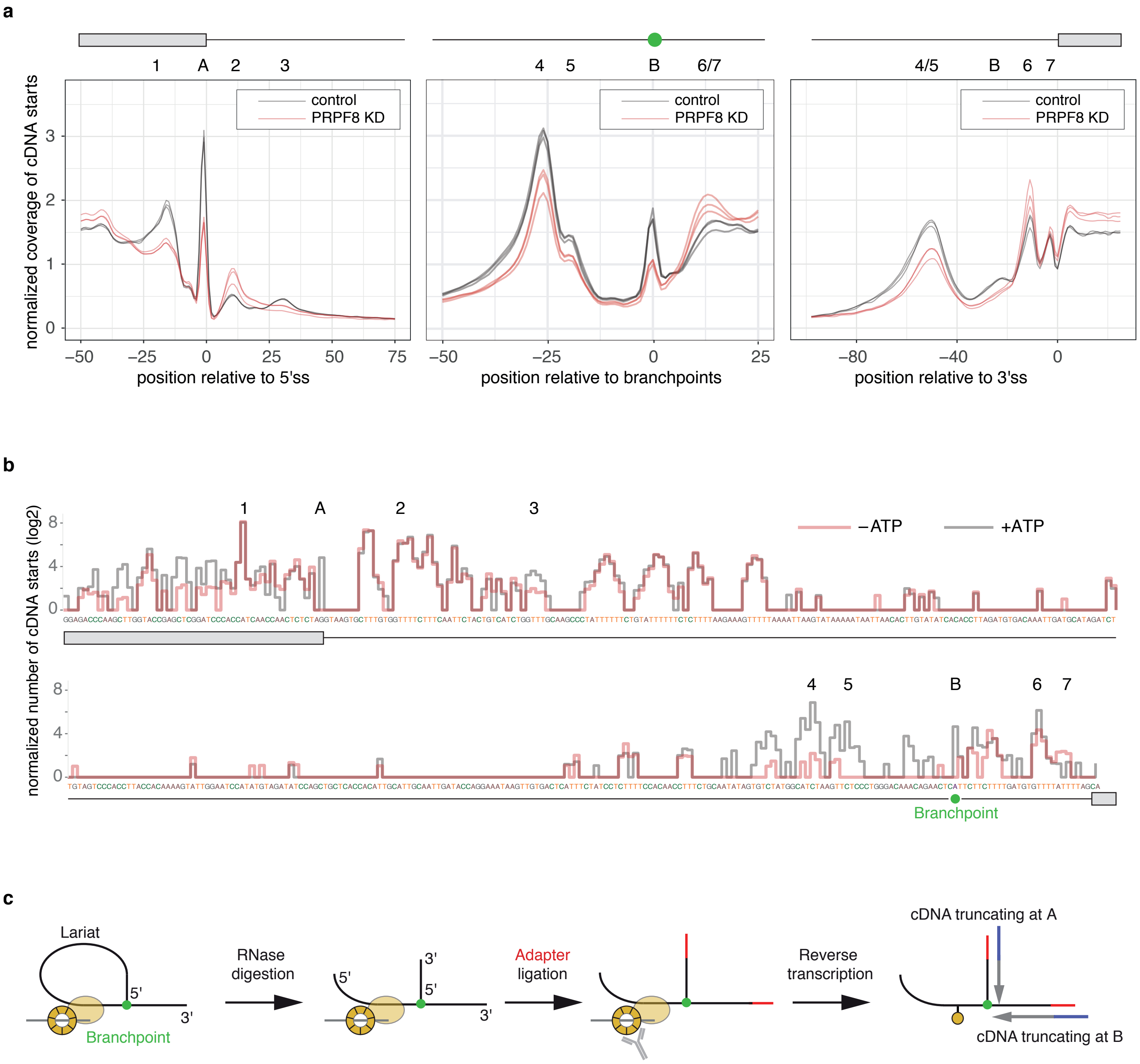

a
snRNPs
UV
UV
UV
UV
D3
B
G
D1
E
F
D2
snRNA
pre-mRNA
in vivo UV crosslinking
stringent
medium
mild
Partial RNase digestion
Immunoprecipitation
Ligation of the 3' adapter
5'
3'
snRNA
3'
3'
5'
snRNA
pre-mRNA
continue
to iCLIP
3'
3'
5'
snRNA
pre-mRNA
b
no Ab
a-SmB/B'
++
+
++
+
RNase I
kDa
185
115
80
65
50
30
25
15
SmB/B'
c
stringent
medium
mild
Percentage of crosslinks (%)
100
75
50
25
0
snRNA
intron
UTR
CDS
lncRNA

Figure 2



a
1
A
2
3
4
5
B
6/7
4/5
B
6
7
control
PRPF8 KD
normalized coverage of cDNA starts
0
1
2
3
−50
−25
0
25
50
75
position relative to 5'ss
−50
−25
0
25
position relative to branchpoints
−80
−40
0
position relative to 3'ss
b
1
A
2
3
−ATP
+ATP
normalized number of cDNA starts (log2)
8
4
0
GGAGACCCAAGCTTGGTACCGAGCTCGGATCCCACCATCAACCAACTCTCTAGGTAAGTGCTTTGTGGTTTTCTTTCAATTCTACTGTCATCTGGTTTGCAAGCCCTATTTTTTCTGTATTTTTTCTCTTTTAAGAAAGTTTTTAAAATTAAGTATAAAAATAATTAACACTTGTATATCACACCTTAGATGTGACAAATTGATGCATAGATCT
4
5
B
6
7
TGTAGTCCCACCTTACCACAAAAGTATTGGAATCCATATGTAGATATCCAGCTGCTCACCACATTGCATTGCAATTGATACCAGGAAATAAGTTGTGACTCATTTCTATCCTCTTTTCCACAACCTTTCTGCAATATAGTGTCTATGGCATCTAAGTTCTCCCTGGGACAAACAGAACTCATTCTTCTTTTGATGTGTTTTATATTTTAGCA
Branchpoint
c
Lariat
5'
3'
Branchpoint
RNase
digestion
5'
3'
5'
3'
Adapter
ligation
Reverse
transcription
cDNA truncating at A
cDNA truncating at B

Figure 3

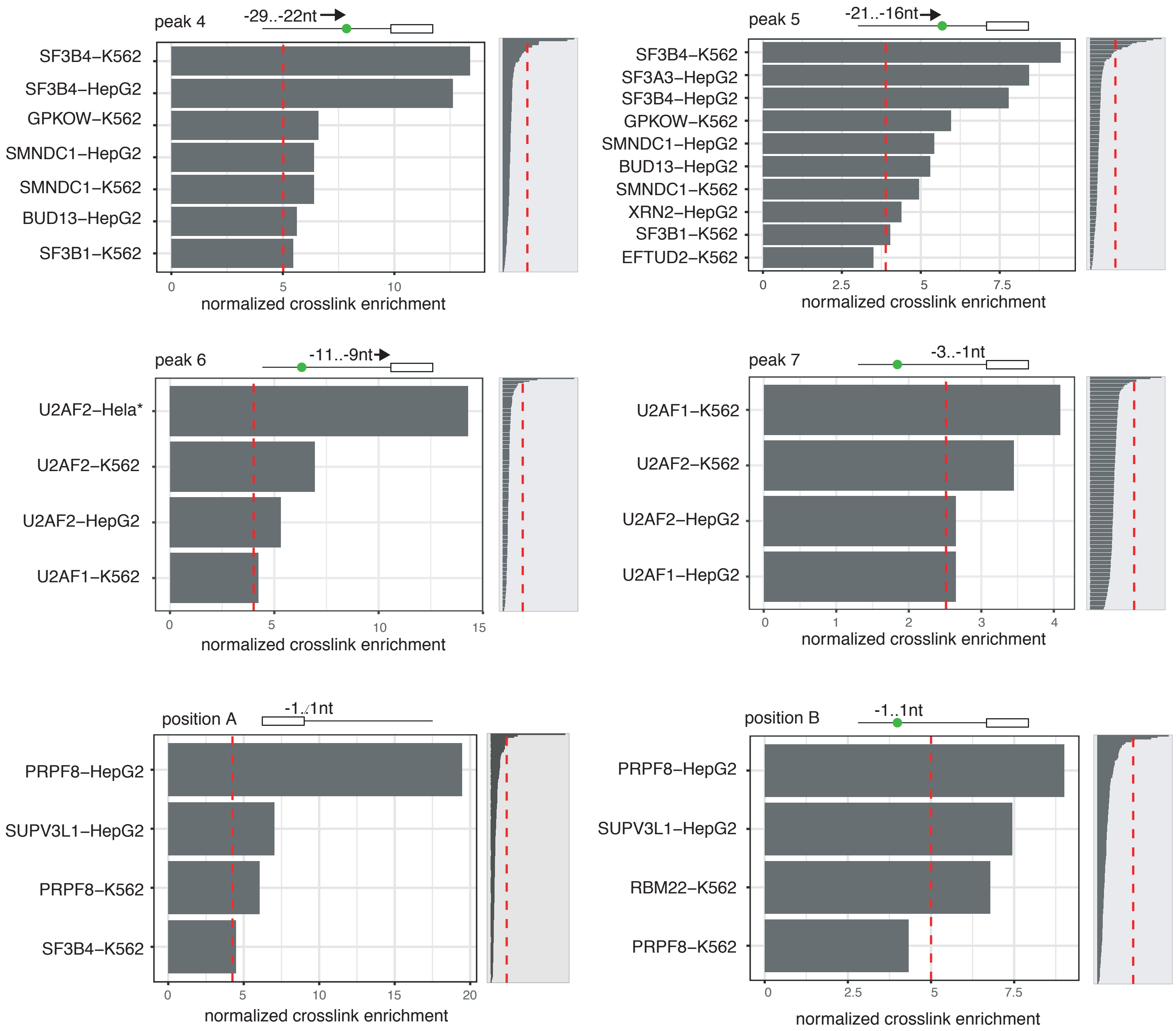
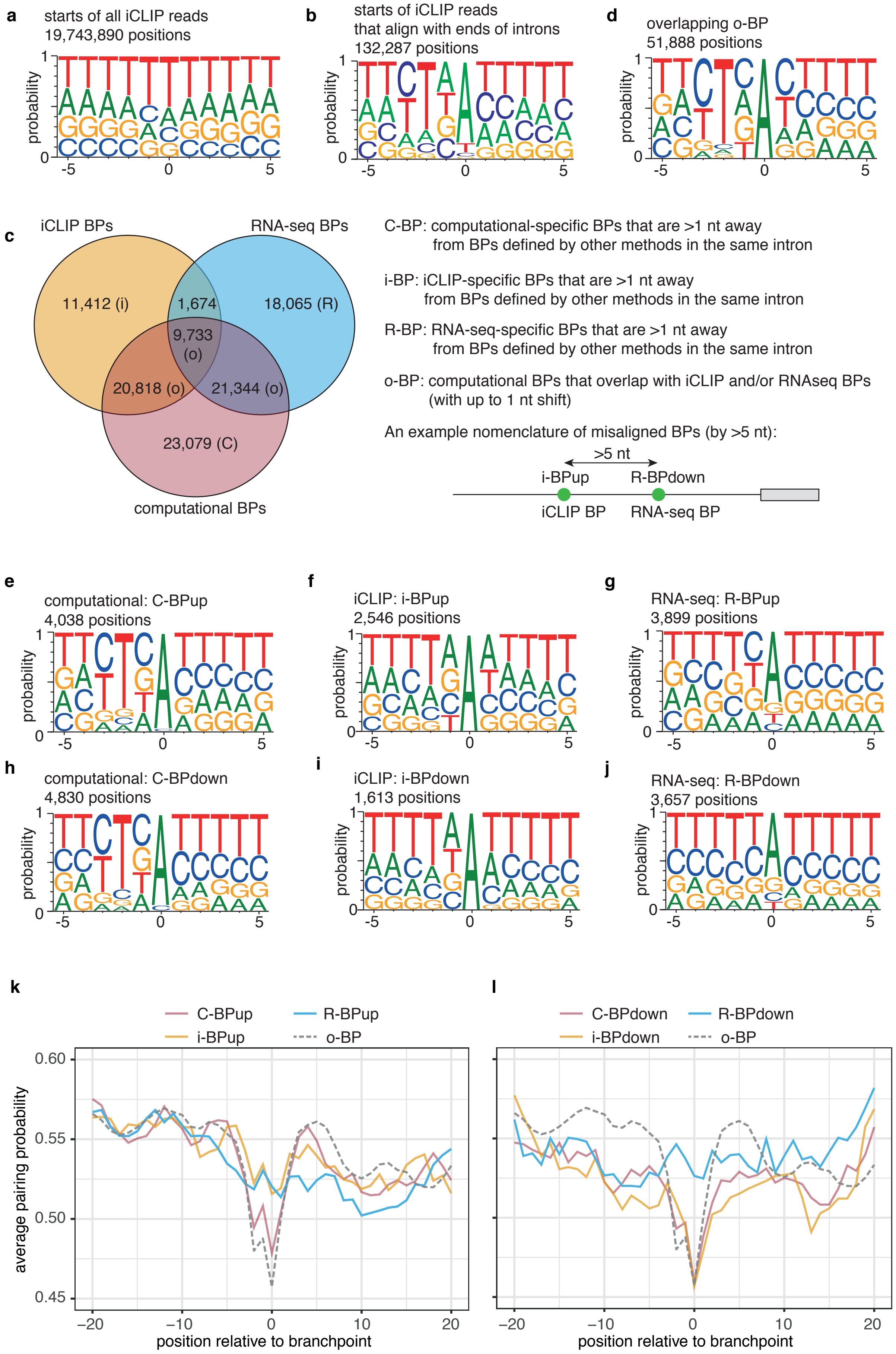

peak 4
-29..-22nt
SF3B4–K562
SF3B4–HepG2
GPKOW–K562
SMNDC1–HepG2
SMNDC1–K562
BUD13–HepG2
SF3B1–K562
0
5
10
normalized crosslink enrichment
peak 5
-21..-16nt
SF3B4–K562
SF3A3–HepG2
SF3B4–HepG2
GPKOW–K562
SMNDC1–HepG2
BUD13–HepG2
SMNDC1–K562
XRN2–HepG2
SF3B1–K562
EFTUD2–K562
0
2.5
5
7.5
normalized crosslink enrichment
peak 6
-11..-9nt
U2AF2–Hela*
U2AF2–K562
U2AF2–HepG2
U2AF1–K562
0
5
10
15
normalized crosslink enrichment
peak 7
-3..-1nt
U2AF1–K562
U2AF2–K562
U2AF2–HepG2
U2AF1–HepG2
0
1
2
3
4
normalized crosslink enrichment
position A
-1..1nt
PRPF8–HepG2
SUPV3L1–HepG2
PRPF8–K562
SF3B4–K562
0
5
10
15
20
normalized crosslink enrichment
position B
-1..1nt
PRPF8–HepG2
SUPV3L1–HepG2
RBM22–K562
PRPF8–K562
0
2.5
5
7.5
normalized crosslink enrichment

Figure 4



C-BP: computational-specific BPs that are >1 nt away from BPs defined by other methods in the same intron

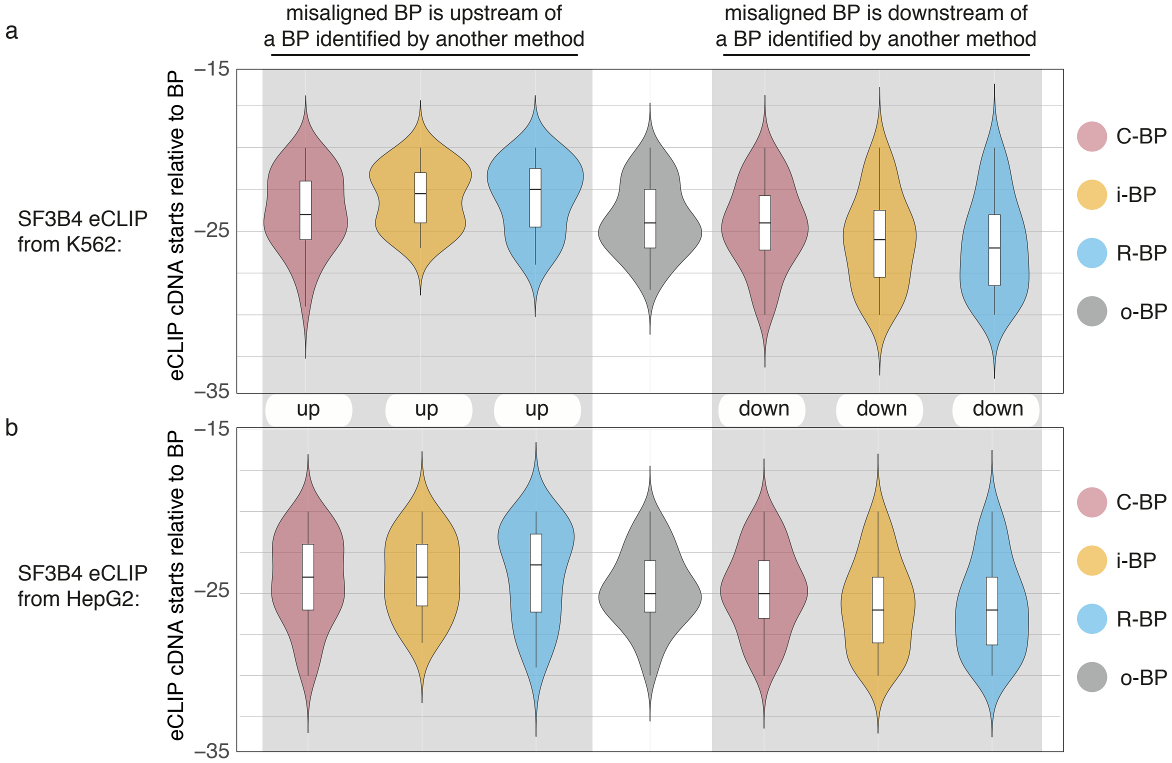
i-BP: iCLIP-specific BPs that are >1 nt away from BPs defined by other methods in the same intron

R-BP: RNA-seq-specific BPs that are >1 nt away from BPs defined by other methods in the same intron

o-BP: computational BPs that overlap with iCLIP and/or RNAseq BPs (with up to 1 nt shift)

An example nomenclature of misaligned BPs (by >5 nt):

Figure 5



BP count:	4038	2546	3899	51895	4,830	1,613	3657
SF3B4 eCLIP, K562:	16078	4947	9791	239162	10276	38204	34116
SF3B4 eCLIP, HepG2:	13545	5396	7326	192818	8248	7556	4903

Figure 6

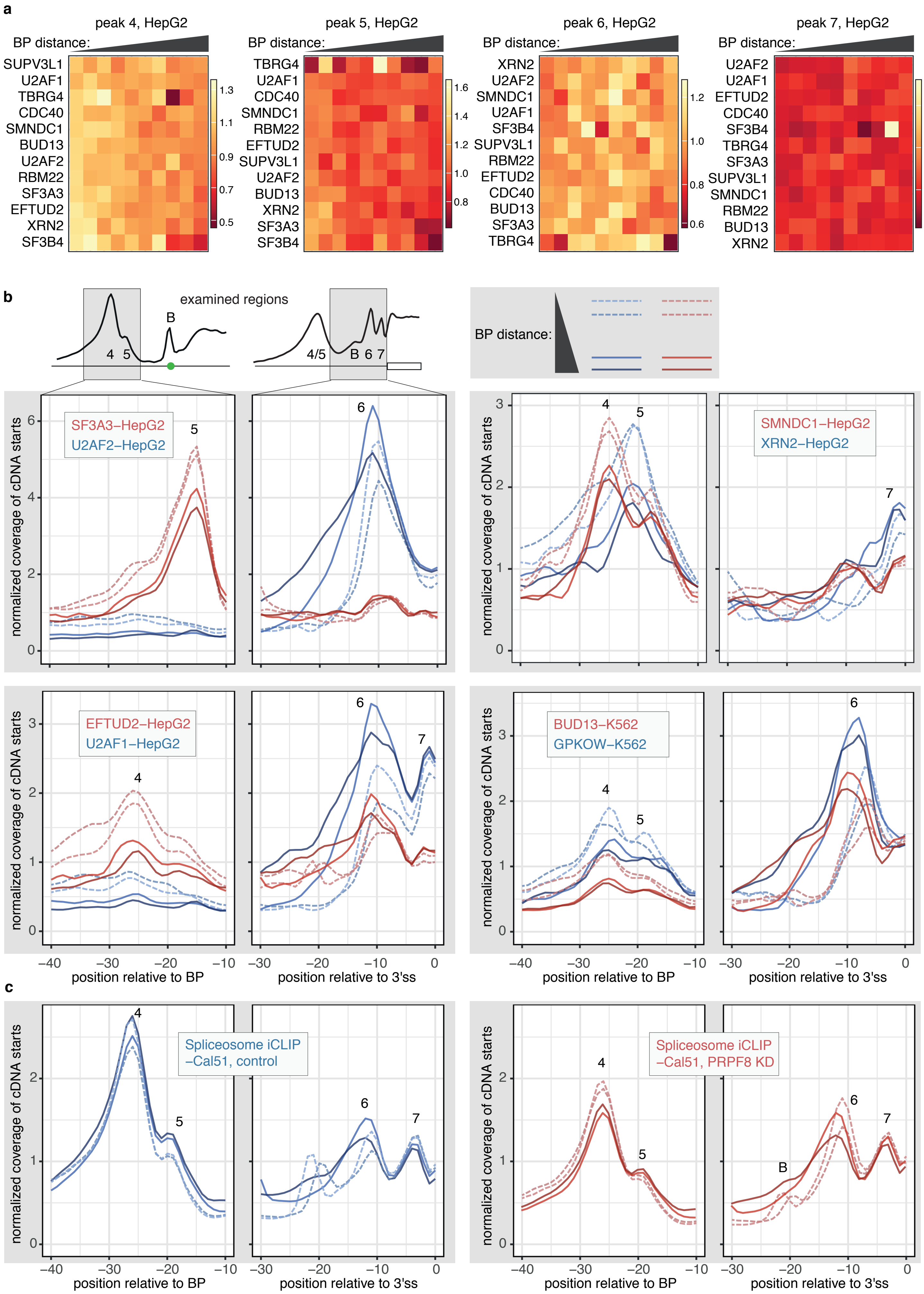
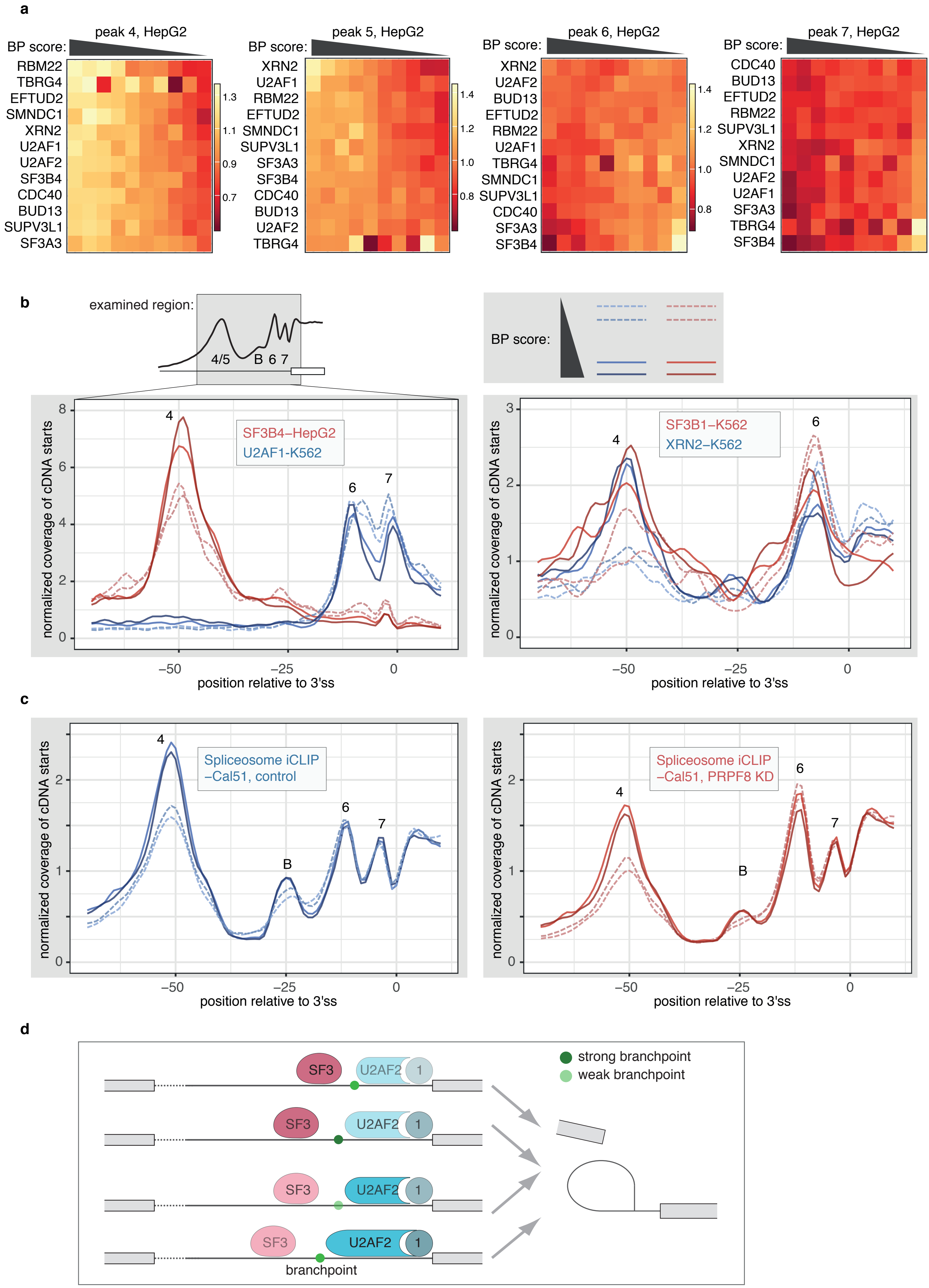

a
peak 4, HepG2
BP distance:
SUPV3L1
U2AF1
TBRG4
CDC40
SMNDC1
BUD13
U2AF2
RBM22
SF3A3
EFTUD2
XRN2
SF3B4
1.3
1.1
0.9
0.7
0.5
peak 5, HepG2
BP distance:
TBRG4
U2AF1
CDC40
SMNDC1
RBM22
EFTUD2
SUPV3L1
U2AF2
BUD13
XRN2
SF3A3
SF3B4
1.6
1.4
1.2
1.0
0.8
peak 6, HepG2
BP distance:
XRN2
U2AF2
SMNDC1
U2AF1
SF3B4
SUPV3L1
RBM22
EFTUD2
CDC40
BUD13
SF3A3
TBRG4
1.2
1.0
0.8
0.6
peak 7, HepG2
BP distance:
U2AF2
U2AF1
EFTUD2
CDC40
SF3B4
TBRG4
SF3A3
SUPV3L1
SMNDC1
RBM22
BUD13
XRN2
2.0
1.5
1.0
b
examined regions
B
4 5
4/5
B 6 7
BP distance:
SF3A3–HepG2
U2AF2–HepG2
SMNDC1–HepG2
XRN2–HepG2
EFTUD2–HepG2
U2AF1–HepG2
BUD13–K562
GPKOW–K562
normalized coverage of cDNA starts
4
5
6
7
position relative to BP
position relative to 3'ss
−40
−30
−20
−10
−30
−20
−10
0
c
Spliceosome iCLIP
–Cal51, control
Spliceosome iCLIP
–Cal51, PRPF8 KD
normalized coverage of cDNA starts
4
5
6
7
B
position relative to BP
position relative to 3'ss

Figure 7


a
peak 4, HepG2
BP score:
RBM22
TBRG4
EFTUD2
SMNDC1
XRN2
U2AF1
U2AF2
SF3B4
CDC40
BUD13
SUPV3L1
SF3A3
1.3
1.1
0.9
0.7
peak 5, HepG2
BP score:
XRN2
U2AF1
RBM22
EFTUD2
SMNDC1
SUPV3L1
SF3A3
SF3B4
CDC40
BUD13
U2AF2
TBRG4
1.4
1.2
1.0
0.8
peak 6, HepG2
BP score:
XRN2
U2AF2
BUD13
EFTUD2
RBM22
U2AF1
TBRG4
SMNDC1
SUPV3L1
CDC40
SF3A3
SF3B4
1.4
1.2
1.0
0.8
peak 7, HepG2
BP score:
CDC40
BUD13
EFTUD2
RBM22
SUPV3L1
XRN2
SMNDC1
U2AF2
U2AF1
SF3A3
TBRG4
SF3B4
1.4
1.2
1.0
0.8
b
examined region:
4/5
B
6
7
BP score:
normalized coverage of cDNA starts
8
6
4
2
0
4
6
7
SF3B4−HepG2
U2AF1−K562
−50
−25
0
position relative to 3'ss
normalized coverage of cDNA starts
3
2
1
0
4
6
SF3B1−K562
XRN2−K562
−50
−25
0
position relative to 3'ss
c
normalized coverage of cDNA starts
2
1
0
4
B
6
7
Spliceosome iCLIP
−Cal51, control
−50
−25
0
position relative to 3'ss
normalized coverage of cDNA starts
2
1
0
4
B
6
7
Spliceosome iCLIP
−Cal51, PRPF8 KD
−50
−25
0
position relative to 3'ss
d
SF3
U2AF2
1
SF3
U2AF2
1
SF3
U2AF2
1
SF3
U2AF2
1
branchpoint
strong branchpoint
weak branchpoint