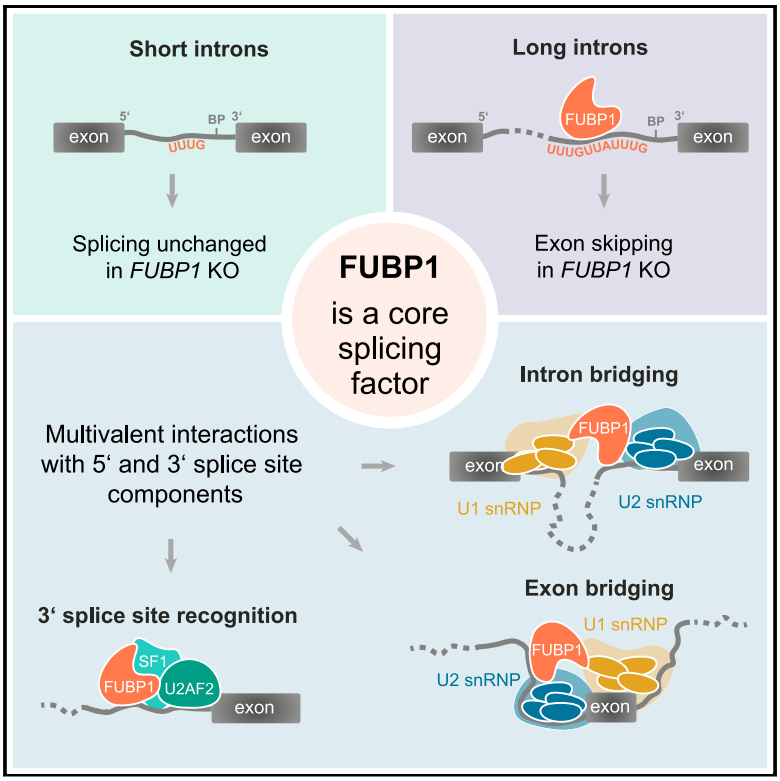
Molecular Cell

Article

FUBP1 is a general splicing factor facilitating 3′ splice site recognition and splicing of long introns

Graphical abstract



Highlights

- FUBP1 recognizes a ubiquitous *cis*-regulatory RNA motif upstream of the branch point
- Multivalent interactions in disordered FUBP1 regions support spliceosome assembly
- FUBP1 affects long introns, which are prevalent in humans and altered in cancer
- Kinetic modeling and protein interactions implicate FUBP1 in splice site bridging

Authors

Stefanie Ebersberger, Clara Hipp, Miriam M. Mulorz, ..., Katja Luck, Michael Sattler, Julian König

Correspondence

k.luck@imb-mainz.de (K.L.), michael.sattler@helmholtz-munich.de (M.S.), j.koenig@imb-mainz.de (J.K.)

In brief

Ebersberger et al. identify the RNA-binding protein FUBP1 as a key splicing factor that binds to a hitherto unknown *cis*-regulatory motif at 3′ splice sites. Multivalent interactions of FUBP1 with splice site components support spliceosome assembly at multiple stages and ensure efficient splicing of long introns.

Ebersberger et al., 2023, Molecular Cell *83*, 2653–2672
August 3, 2023
https://doi.org/10.1016/j.molcel.2023.07.002

CellPress

Molecular Cell

CellPress
OPEN ACCESS

Article

FUBP1 is a general splicing factor facilitating 3′ splice site recognition and splicing of long introns

Stefanie Ebersberger,[1,12] Clara Hipp,[2,3,12] Miriam M. Mulorz,[1,12] Andreas Buchbender,[1] Dalmira Hubrich,[1] Hyun-Seo Kang,[2,3] Santiago Martínez-Lumbreras,[2,3] Panajot Kristofori,[4] F.X. Reymond Sutandy,[1] Lidia Llacsahuanga Allcca,[1,13] Jonas Schönfeld,[1] Cem Bakisoglu,[5] Anke Busch,[1] Heike Hänel,[1] Kerstin Tretow,[1] Mareen Welzel,[1] Antonella Di Liddo,[1] Martin M. Möckel,[1] Kathi Zarnack,[5,6] Ingo Ebersberger,[7,8,9] Stefan Legewie,[10,11] Katja Luck,[1,*] Michael Sattler,[2,3,*] and Julian König[1,14,*]

[1]Institute of Molecular Biology (IMB) gGmbH, 55128 Mainz, Germany
[2]Institute of Structural Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany
[3]Bavarian NMR Center, Department of Bioscience, School of Natural Sciences, Technical University of Munich, 85747 Garching, Germany
[4]Department of Systems Biology, Institute for Biomedical Genetics (IBMG), University of Stuttgart, 70569 Stuttgart, Germany
[5]Buchmann Institute for Molecular Life Sciences & Institute of Molecular Biosciences, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany
[6]CardioPulmonary Institute (CPI), 35392 Gießen, Germany
[7]Applied Bioinformatics Group, Institute of Cell Biology and Neuroscience, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany
[8]Senckenberg Biodiversity and Climate Research Center (S-BIK-F), 60325 Frankfurt am Main, Germany
[9]LOEWE Center for Translational Biodiversity Genomics (TBG), 60325 Frankfurt am Main, Germany
[10]Department of Systems Biology, Institute for Biomedical Genetics (IBMG), University of Stuttgart, 70569 Stuttgart, Germany
[11]Stuttgart Research Center for Systems Biology (SRCSB), University of Stuttgart, 70569 Stuttgart, Germany
[12]These authors contributed equally
[13]Present address: University of California, Berkeley, CA 94720, USA
[14]Lead contact
*Correspondence: k.luck@imb-mainz.de (K.L.), michael.sattler@helmholtz-munich.de (M.S.), j.koenig@imb-mainz.de (J.K.)
https://doi.org/10.1016/j.molcel.2023.07.002

SUMMARY

Splicing of pre-mRNAs critically contributes to gene regulation and proteome expansion in eukaryotes, but our understanding of the recognition and pairing of splice sites during spliceosome assembly lacks detail. Here, we identify the multidomain RNA-binding protein FUBP1 as a key splicing factor that binds to a hitherto unknown *cis*-regulatory motif. By collecting NMR, structural, and *in vivo* interaction data, we demonstrate that FUBP1 stabilizes U2AF2 and SF1, key components at the 3′ splice site, through multivalent binding interfaces located within its disordered regions. Transcriptional profiling and kinetic modeling reveal that FUBP1 is required for efficient splicing of long introns, which is impaired in cancer patients harboring *FUBP1* mutations. Notably, FUBP1 interacts with numerous U1 snRNP-associated proteins, suggesting a unique role for FUBP1 in splice site bridging for long introns. We propose a compelling model for 3′ splice site recognition of long introns, which represent 80% of all human introns.

INTRODUCTION

Splicing is a crucial step in eukaryotic mRNA processing, and its dysregulation is a hallmark of many cancers.[1–3] Splicing is catalyzed by the spliceosome, a megadalton machinery comprising five small nuclear ribonucleoprotein (snRNP) complexes named U1, U2, U4, U5, and U6.[4–7] During early spliceosome assembly (E complex formation), the 5′ and 3′ splice sites are recognized: U1 binds at the 5′ splice site, whereas U2 auxiliary factor 1 (U2AF1), U2AF2, and splicing factor 1 (SF1) assemble at the 3′ splice site,[6–11] where they specifically recognize AG dinucleotide,[12,13] polypyrimidine (Py) tract,[14–16] and branch point (BP) site, respectively (Figure 1A).[9,17] In the resulting A complex, U2 snRNP is recruited to the BP and stabilized by SF3A and SF3B, and SF1 is released.[18,19] Subsequent snRNP recruitment and further rearrangements (formation of B and C complexes) mediate intron excision and exon ligation to form the mature mRNA.

Strikingly, mechanistic details of splice site recognition by multidomain splicing factors during early spliceosome assembly are lacking.[20,21] U2AF2 binding is central to the early definition of splice sites and is subject to layers of regulation including direct

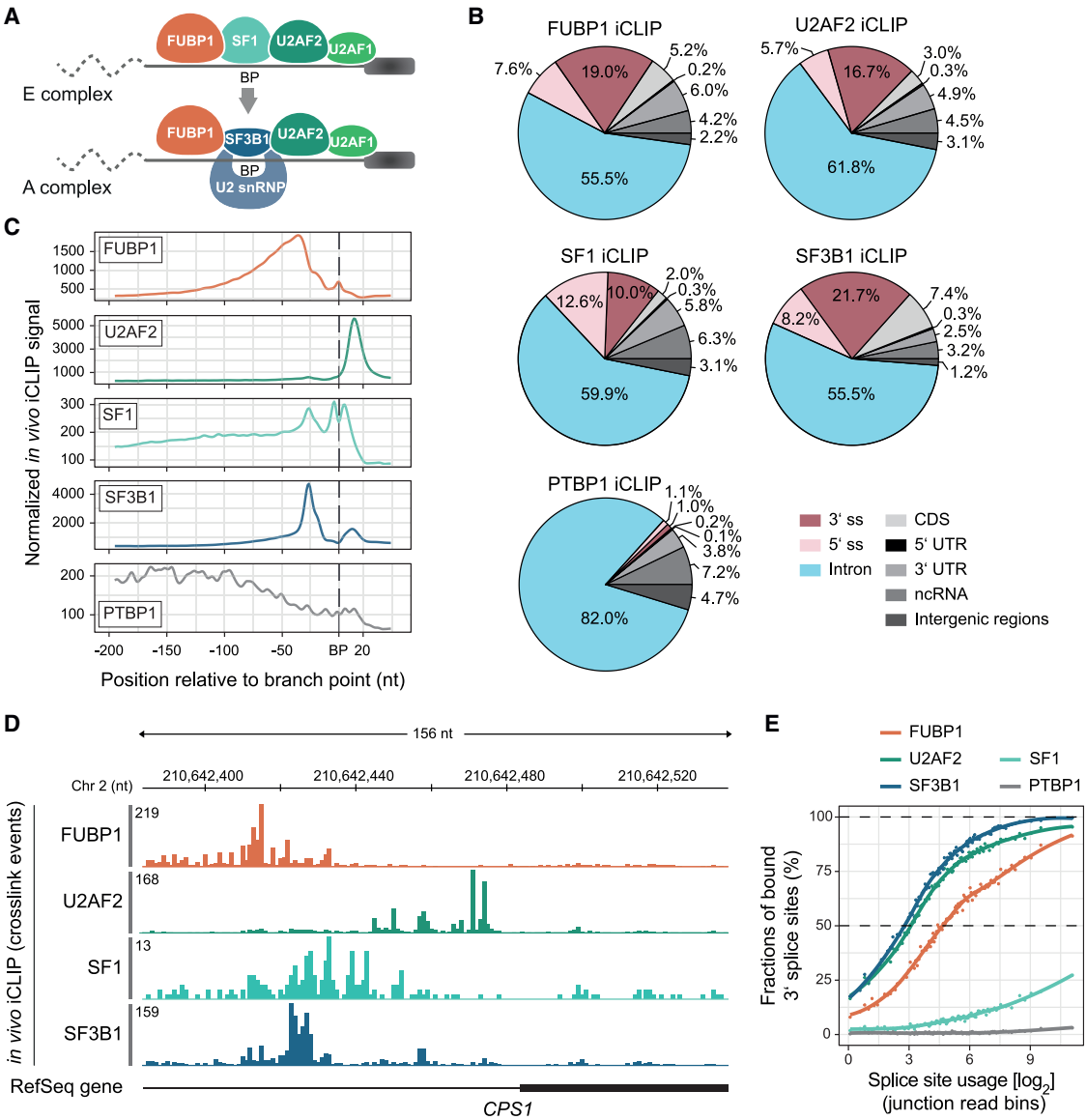


Figure 1. FUBP1 binds upstream of the branch point at 3′ splice sites during early spliceosome assembly *in vivo*

(A) Schematic of spatial RBP assembly at the 3′ splice site in the "commitment" E complex and the pre-spliceosomal A complex. BP, branch point.

(B) iCLIP in HeLa cells. Distribution of binding sites across transcript regions for FUBP1 (n = 854,404), U2AF2 (n = 914,221), SF1 (n = 99,305), SF3B1 (n = 1,694,991), and PTBP1 (n = 127,450). 3′ and 5′ splice sites (ss) refer to 100 nt upstream/downstream of exons, respectively. CDS, coding sequence; UTR, untranslated region.

(C) Metaprofiles of cross-link events of FUBP1, U2AF2, SF1, SF3B1, and PTBP1 relative to the BP.

(D) Genome browser view of an internal exon in the *CPS1* mRNA displaying the iCLIP data for FUBP1, U2AF2, SF1, and SF3B1 from HeLa cells.

(E) Saturation analysis showing the percentage of bound 3′ splice sites for each RBP in each quantile.

competition, cooperative recruitment, change of RNA secondary structure, dynamic conformational states, and autoinhibition.[15,22–30] Despite the pivotal role of U2AF2, the precise contribution of cofactors and multivalent interactions are yet to be elucidated. Recently, we reported how U2AF2 achieves specificity despite the degeneracy of its pyrimidine-rich RNA-binding motif.[28] In this study, we found that the RNA-binding protein (RBP) far upstream binding protein 1 (FUBP1) promotes U2AF2 binding to RNA.

FUBP1 was initially characterized as a transcriptional regulator of the proto-oncogene *c-myc* through binding to AT-rich DNA elements and interaction with PUF60, also known as the FUBP-interacting repressor (FIR).[31–34] However, more recently, FUBP1 has also been reported to bind RNA and to influence translation or splicing of specific transcripts.[35–38] Similar to its DNA-binding specificity, FUBP1 exhibits a general preference for AU- and GU-rich RNA[31] that is expected to derive from its four K homology (KH) domains.[39] Notably, cancer-associated

CellPress
OPEN ACCESS

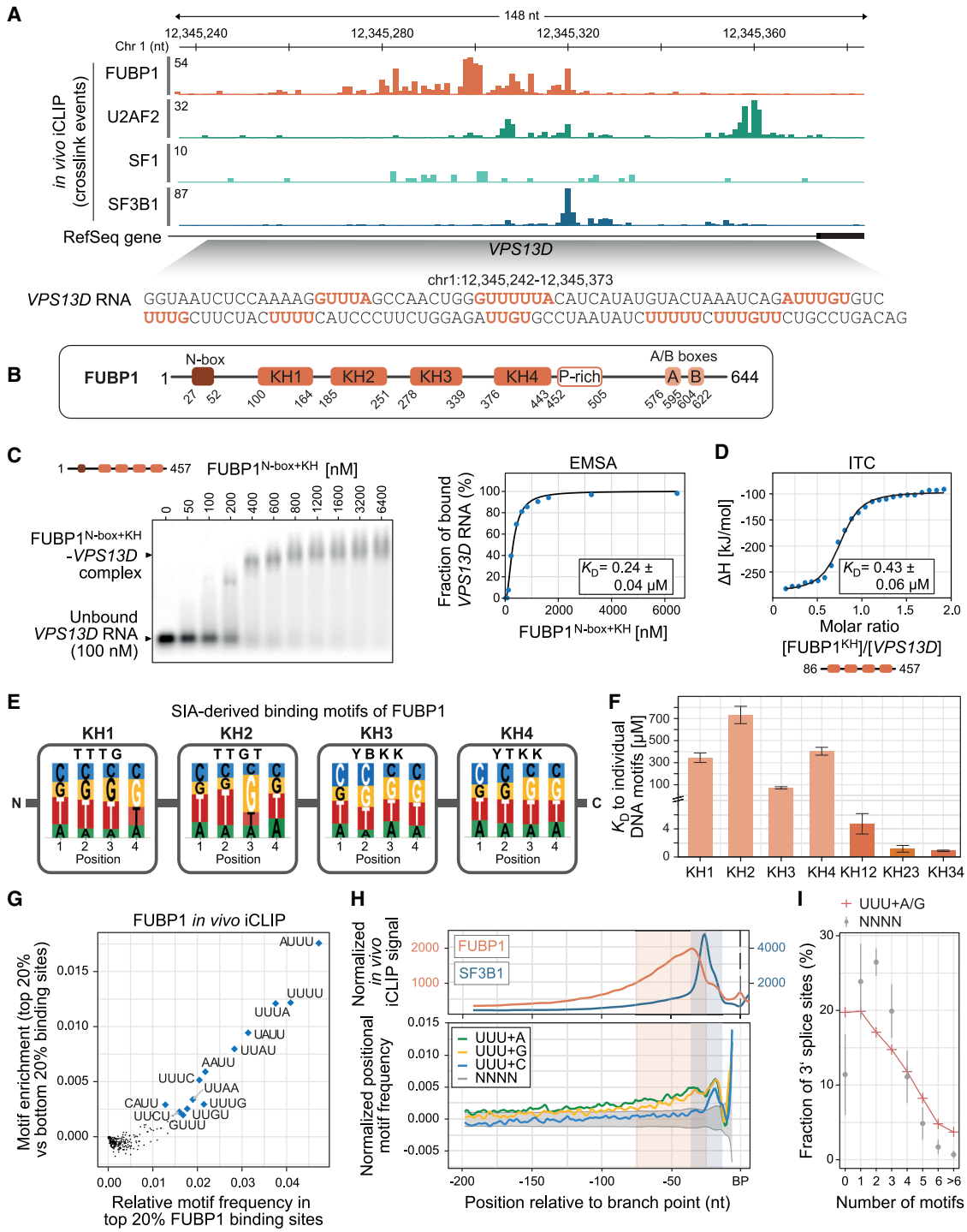


Figure 2. FUBP1 binds a hitherto unknown *cis*-regulatory motif upstream of the BP

(A) Genome browser view of an internal exon in the *VPS13D* mRNA displaying the iCLIP data for FUBP1, U2AF2, SF1, and SF3B1.

(B) Domain architecture of FUBP1. KH, K homology domain; P-rich, proline-rich stretch.

(C) Agarose gel (left) and quantification with fitted curve (right) from an EMSA experiment with recombinant FUBP1$^{N\text{-box+KH}}$ (50–6,400 nM) and a fluorescently labeled 132-nt RNA fragment of *VPS13D* (100 nM). Measurements were performed in duplicates and data are represented as mean ± standard deviation (SD).

(D) Binding affinity for the interaction of FUBP1^{KH} with *VPS13D* RNA determined by ITC. ITC measurements were performed in triplicates and data are represented as mean ± SD.

(legend continued on next page)

loss-of-function mutations within *FUBP1* have been connected to global splicing changes in low-grade glioma,[1,40–42] suggesting an RNA-regulatory role in these processes. Here, we reveal a global role for FUBP1 in pre-mRNA splicing. Our results suggest that FUBP1 functions as a general splicing factor at the 3′ splice site, with a crucial role in promoting efficient splicing of long introns, which make up over 80% of human pre-mRNA transcripts.

RESULTS

FUBP1 is a core component of 3′ splice site recognition

To dissect the role of FUBP1 in splicing, we examined the footprint of FUBP1 and other splicing factors on pre-mRNA in HeLa cells using *in vivo* individual-nucleotide resolution UV cross-linking and immunoprecipitation (*in vivo* iCLIP; Figures 1B and S1A; Table S1).[43,44] As expected, large proportions of the binding sites of SF1, U2AF2, and SF3B1 are located at 3′ splice sites (10%, 17%, and 22%, respectively). Interestingly, FUBP1 shows a similar preference for 3′ splice sites (19%). By contrast, for the more restricted splicing regulator PTBP1, which is known to act on a subset of exons, only 1% of binding sites are located at 3′ splice sites. We confirmed that U2AF2 binds at the Py tract located between the BP and 3′ splice site,[45,46] whereas SF1 binding peaks at the BP, with a reduced signal at the BP adenine itself,[9,17] presumably owing to the lower cross-linking efficiency of adenine (Figures 1C and 1D).[47] Consistent with a previous report,[48] SF3B1 binds in a clamp-wise manner up- and downstream of the BP. Strikingly, FUBP1 also shows a pronounced footprint at the BP (Figures 1C and 1D). Its binding peaks at a location 34 nucleotides (nt) upstream of the BP and tails for up to 100 nt. In comparison, PTBP1 does not display such a ubiquitous positioning at 3′ splice sites (Figure 1C).[49,50] Next, we addressed what fraction of 3′ splice sites is bound using a saturation-based analysis that controls for splice site usage and transcript abundance.[51] We found that FUBP1 binds the same percentage of 3′ splice sites as U2AF2 and SF3B1, which are both universally present at 3′ splice sites (91.3%, 95.4%, and 99.6%, respectively; Figure 1E). By contrast, SF1 and PTBP1 are associated with 27.3% and 3.1% of 3′ splice sites, respectively (Figures 1E and S1B). Overall, these data suggest that FUBP1 functions as a general splicing factor in early spliceosome assembly.

FUBP1 binds a *cis*-regulatory RNA motif upstream of the branch point

Given the prevalence of FUBP1 upstream of the BP, we investigated its RNA-binding preferences. First, we performed electrophoretic mobility shift assays (EMSAs) with a 132-nt RNA fragment upstream of the prototypical 3′ splice site of exon 43 of the *VPS13D* mRNA (*VPS13D*) and a shortened fragment (36 nt) with the region showing the most FUBP1 binding in iCLIP (*VPS13D*^{short}; Figure 2A). We observed strong binding of FUBP1 (FUBP1$^{N\text{-}box+KH}$, aa 1–457) to both RNAs in the low nanomolar range (Figures 2B, 2C, and S1C). Isothermal titration calorimetry (ITC) with *VPS13D* yielded a similar result (Figure 2D; Table S2), confirming the high-affinity binding at this region.

FUBP1 harbors four KH domains, which are expected to bind single-stranded RNA and DNA[32,52] and can act either independently or synergistically[53–55] to recognize extended regions of pre-mRNA. We used nuclear magnetic resonance (NMR) spectroscopy to investigate the modular arrangement of the four FUBP1 KH domains. Superimposition showed that the NMR spectrum of FUBP1^{KH} (aa 86–457) containing KH1–4 was virtually identical to those of the individual KH domains, indicating that the KH domains are structurally independent (Figure S1D). Furthermore, NMR secondary structure analysis revealed that FUBP1 contains KH domains with a typical type I fold that are connected by flexible linkers (Figure S1E).[56] We conclude that the KH domains of FUBP1 are not preformed into an RNA-binding platform but rather can be considered like beads on a string.

To characterize the individual RNA-binding preferences of the four KH domains, we performed a scaffold-independent analysis (SIA), which is based on changes in NMR chemical shifts upon titration with short oligonucleotide motifs (Figure S2A).[57] Initial binding experiments were performed using randomized pools of 5-mer DNA, followed by verification of the identified motifs using RNA oligonucleotides (Figure S2B). SIA identified well-defined consensus motifs for KH1 (UUUG) and KH2 (UUGU) and more loosely defined motifs for KH3 (YBKK, where Y = C or U; B = C, G, or U; K = G or U) and KH4 (YUKK). Hence, all four KH domains exhibit a preference for GU-rich sequences (Figure 2E). The affinities of the individual KH domains to the final motifs, as determined by NMR spectroscopy, are in the high micromolar range (Figures S2C–S2F). Combinations of two KH domains and motifs show strong binding avidity: the ITC-measured affinities for tandem domains were in the high nanomolar to low micromolar range (Figures S2G–S2I; Table S2). This suggests that specificity and high affinity are achieved by avidity and multivalent interactions between the four KH domains and RNA with multiple binding motifs (Figure 2F). Indeed, EMSA and ITC experiments confirmed that multiple FUBP1 binding motifs in the *VPS13D* mRNA fragment increase FUBP1 binding to nanomolar affinity (Figures 2A, 2C, and 2D).

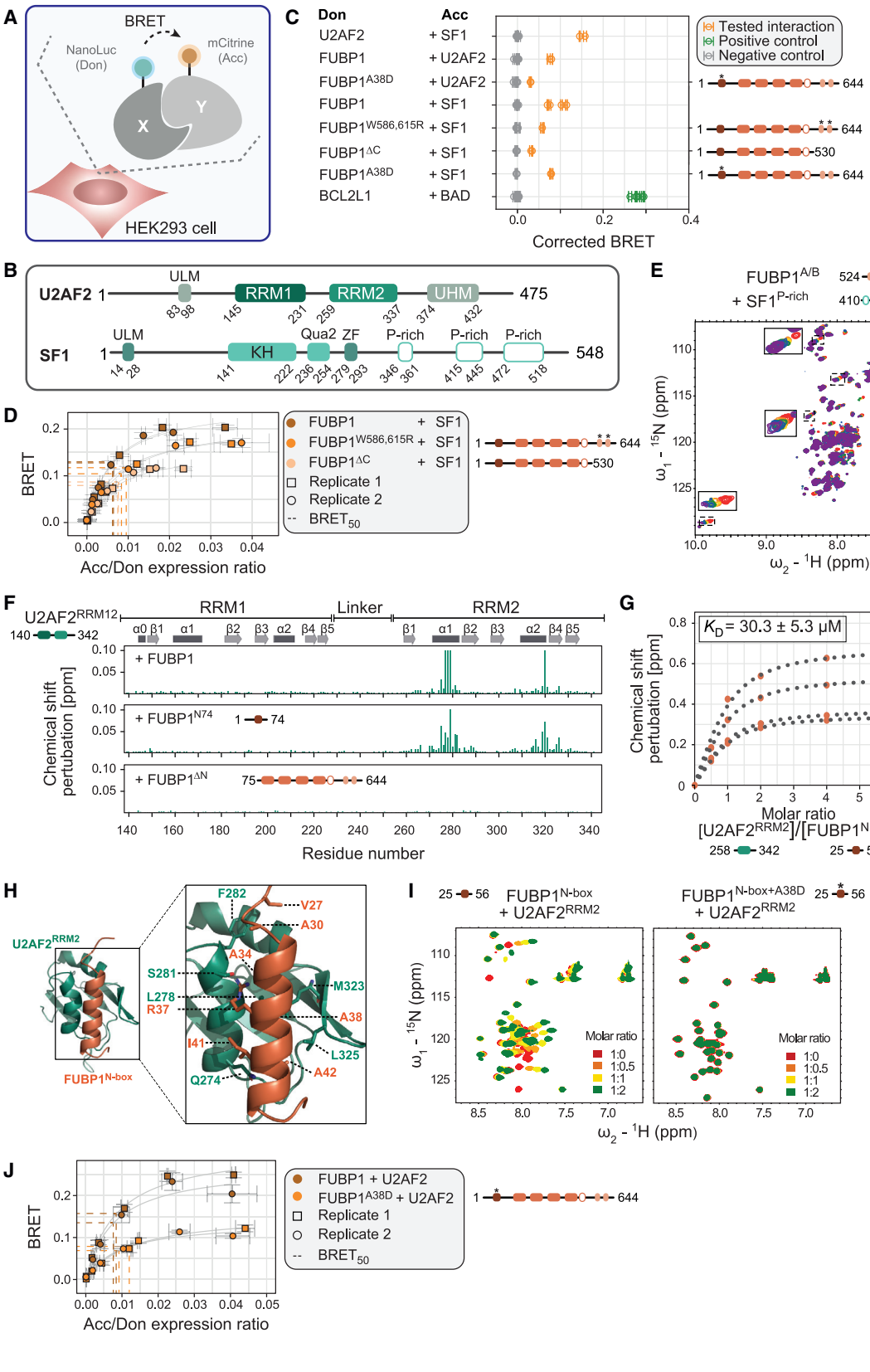
(E) Scaffold-independent analysis (SIA)-derived binding motifs for individual FUBP1 KH domains. Preferred bases are highlighted in white. Y, pyrimidine (T or C); B, not A (C, G, or T); K, keto (G or T).

(F) K_D values of individual and tandem KH domains with their optimal DNA target (KH1, TTTTG; KH2, TTTGT; KH3, TCTGT; KH4, TTTTG; KH1-2, TTTGTAAAATTTTG; KH2-3, TCTGTAAAATTTGT; KH3-4, TTTTGAAAATCTGT) determined by NMR or ITC, respectively (Figures S2C–S2I; Table S2). ITC measurements were performed in triplicates. For NMR, the K_D values of eight selected residues were calculated. Data are represented as mean ± SD.

(G) Motif enrichment in the *in vivo* FUBP1 iCLIP data. Disjunct 4-mer frequencies were calculated for the top vs bottom 20% of binding sites based on expression-normalized iCLIP signals.

(H) Positional enrichment of FUBP1 binding motifs and control motifs relative to the BP. UUU+A/G/C, i.e., 4-mers containing UUU interspersed at any position with A/G/C. NNNN, 100,000 sets of random combinations of four 4-mers. 4-mer frequencies were calculated position-wise upstream of the BP and compared with the average 4-mer frequencies in an intronic control region. Top: Metaprofile of normalized FUBP1 and SF3B1 iCLIP cross-link events at the same 3′ splice sites is shown for comparison.

(I) Abundance of FUBP1 binding motifs at 3′ splice sites of human introns. Background distribution for all possible 4-mers (mean ± 1 SD) is shown in gray.



(legend on next page)

Interrupting the U-rich motifs of *VPS13D*^{short} with cytidines severely reduces the binding affinity, underlining the specificity of the FUBP1-RNA interaction (Figure S1C).

To validate the interaction between FUBP1 and RNA motifs in cells, we compared 4-mer motifs in the sites of strongest FUBP1 binding over background in the *in vivo* iCLIP data. In line with the SIA, we found a strong preference for uridine-rich motifs at FUBP1 binding sites (Figures 2G and S2J). For *in vivo* binding, these motifs can be interspersed at any position by adenine or, to a lesser extent, by guanine. Consistent with the omnipresence of FUBP1 at 3′ splice sites, we observed a striking enrichment of FUBP1 binding motifs ("UUU+A" and "UUU+G," i.e., three uridines interspersed at any position with adenine or guanine) upstream of the BP, where they coincide with FUBP1 binding (Figure 2H). Conversely, both "UUU+C," accounting for general uridine richness, and random motif sets are enriched closer to the 3′ splice site but not in the main region of FUBP1 binding. Importantly, enriched FUBP1 motifs upstream of the BP are a common feature across all annotated introns (Figures 2H, 2I, S2K, and S2L), indicating that we identified a previously unknown *cis*-regulatory RNA motif in splicing regulation.

FUBP1 directly interacts with U2AF2 and SF1

Given the prevalence of FUBP1 at functional 3′ splice sites, we examined whether FUBP1 interacts with key early 3′ splice site components in cells using bioluminescence resonance energy transfer (BRET) (Figure 3A).[58] Interaction signals in the BRET assay are indicative of direct contacts or close proximities. As a proof-of-concept, we confirmed the known U2AF2-SF1 interaction.[8,10,18,58] Importantly, we also observed interactions of FUBP1 with U2AF2 and SF1 (Figures 3B, 3C, and S2M), suggesting that FUBP1 is in close or direct contact with these core splicing factors inside cells.

To investigate whether FUBP1 directly interacts with SF1 (Figure 3C), we focused on the C-terminal region of FUBP1, which harbors the A and B boxes (A/B boxes). These motifs are specific to the FUBP family of proteins and have been shown to mediate binding to a proline-rich region of snRNP-U1-70K in fruit flies.[60,59] Similar proline-rich regions are also present in the C-terminal region of SF1. Consistently, the SF1-FUBP1 interaction detected by BRET is reduced upon deletion or mutation of the A/B boxes (Figures 3B–3D and S2M). In addition, ^{1}H-^{15}N correlation NMR spectra of the FUBP1 A/B box region show specific chemical shift perturbations (CSPs) upon titration with the proline-rich region of SF1, indicating direct binding (Figure 3E).

To map the interacting regions in FUBP1 and U2AF2, we performed NMR titration experiments using ^{15}N-labeled U2AF2^{RRM12} [14,15,30] and unlabeled full-length FUBP1. Large CSPs and line broadening in the ^{1}H-^{15}N correlation spectra exclusively map to the U2AF2 RRM2 domain, especially to the two α helices on the backside of the β sheets that mediate RNA binding (Figures 3B, 3F, S2N, and S3A). Moreover, a construct comprising the N-terminal region of FUBP1 (FUBP1^{N74}, aa 1–74) recapitulates the CSPs observed with full-length FUBP1, whereas a construct lacking the N-terminal region (FUBP1$^{\Delta N}$, aa 75–644) does not yield any evident CSPs (Figures 3F and S3A). Complementary NMR titrations with ^{15}N-labeled FUBP1 constructs identify the U2AF2 RRM2 domain and a short peptide motif in the N-terminal region of FUBP1 (aa 27–52), referred to as N-box, as the minimal binding regions (Figures S3B–S3F). The U2AF2 RRM2-FUBP1 N-box interaction exhibits micromolar affinity by NMR titrations (Figures 3G, S3D, S3G, and S3H; Table S2).

To provide a high-resolution view, we determined the NMR-derived solution structure of the U2AF2 RRM2-FUBP1 N-box complex (Figures 3H and S3I–S3K; Table 1). This structure shows a well-defined U2AF2 RRM2 domain and a more mobile helical FUBP1 N-box and reveals that the FUBP1$^{N\text{-box}}$ forms an α helix, which is recognized by helices α1 and α2 and the β4 strand of U2AF2^{RRM2}. Hydrophobic interactions dominate at this interface, where four alanines in FUBP1$^{N\text{-box}}$ (A30, A34, A38, and A42) are aligned along the extended hydrophobic interface, with A38 positioned centrally. Additional contacts involving bulkier side chains, that is, R37 and I41 in FUBP1 and L278 and M323 in U2AF2, further stabilize the binding interface. The recognition of the FUBP1 N-box resembles the interaction between FUBP1 N-box and PUF60,[34] consistent with structural similarities between PUF60 and U2AF2 RRM2 (Figure S4A).[34]

Figure 3. FUBP1 directly interacts with SF1 and U2AF2 via its C-terminal A/B boxes and N-terminal N-box

(A) Schematic of BRET assay. Energy transfer between the substrate oxidized by NanoLuc luciferase (donor, Don) and mCitrine (acceptor, Acc) occurs if proteins X and Y interact.

(B) Domain architecture of U2AF2 (UniProt: P26368) and SF1 (UniProt: Q15637). ULM, U2AF ligand motif; RRM, RNA-recognition motif; UHM, U2AF homology motif family; Qua2, quaking homology 2 domain; ZF, zinc finger.

(C) BRET values for tested interaction pairs and controls. Two biological replicates are shown. Error bars represent SD of technical triplicates. Trp-to-Arg mutations in the A/B boxes were rationalized based on disrupting the hydrophobic contacts as previously reported.[59]

(D) BRET saturation curves for combinations of FUBP1 variants and wild-type SF1. Trp-to-Arg mutations in the A/B boxes or their deletion significantly lowered the maximal BRET signal, although changes in the BRET$_{50}$ (acceptor/donor ratio at which half-maximal BRET signal is reached) were not significant. Amounts of acceptor and donor proteins were estimated by fluorescence and total luminescence, respectively, in intact cells. Two biological replicates are shown. Error bars represent SD of technical triplicates.

(E) NMR titration of FUBP1$^{A/B}$ with SF1$^{P\text{-rich}}$. Significant chemical shift changes are highlighted by boxes.

(F) Binding interface mapping based on NMR titration of U2AF2^{RRM12} with full-length FUBP1, FUBP1^{N74}, and FUBP1$^{\Delta N}$ (Figure S3A).

(G) Binding affinity for the interaction of FUBP1$^{N\text{-box}}$ and U2AF2^{RRM2} from NMR titrations. Chemical shift differences of four exemplary residues of FUBP1$^{N\text{-box}}$ (Figures S3D and S3G) are fitted to binding isotherm to estimate the K_D. Data are represented as mean ± SD of calculated K_D values of eight selected residues.

(H) NMR-derived structure of the complex of U2AF2^{RRM2} (green) and FUBP1$^{N\text{-box}}$ (brown) (Figure S3K; Table 1, PDB: 8P25).

(I) Comparison of NMR titrations of FUBP1$^{N\text{-box}}$ WT and mutant FUBP1$^{N\text{-box+A38D}}$ with U2AF2^{RRM2}.

(J) BRET saturation curves for wild-type FUBP1 and mutant FUBP1^{A38D} against U2AF2. Two biological replicates are shown. Error bars represent SD of technical triplicates.

Table 1. Statistics for structure calculation of the U2AF2^{RRM2}/FUBP1$^{N\text{-}box}$ chimera, related to Figures 3H and S3K, PDB: 8P25[a]

Experimental restraints	
Distance restraints	
Total NOE	2,147
Short range, $\|i–j\| \leq 1$	1,047
Medium range, $1 < \|i–j\| < 5$	392
Long range, $\|i–j\| \geq 5$	708
Dihedral angle restraints (from TALOS)	
Φ	82
Ψ	86
Structure statistics	
RMSD from experimental restraints (mean and SD)	
Distance restraints (Å), no violation > 0.5 Å	0.013 ± 0.007
Dihedral angle restraints (°, no violation > 0.5°)	0.19 ± 0.04
Deviations from idealized geometry	
Bond lengths (Å)	0.004 ± 0.0001
Bond angles (°)	0.60 ± 0.01
Impropers (°)	1.31 ± 0.04
Average pairwise coordinate RMSD (Å)	
Backbone	0.92 ± 0.30
Heavy atoms	1.41 ± 0.22

[a]Pairwise coordinate root-mean-square deviation (RMSD) was calculated for the 10 lowest-energy structures (regions 250–336 in U2AF2^{RRM2} and 31–43 in FUBP1$^{N\text{-}box}$) after water refinement. Ramachandran plot: 93.1%, 6.1%, 0.3%, and 0.4% of residues (regions 250–336 in U2AF2^{RRM2} and 31–43 in FUBP1$^{N\text{-}Box}$) are found in the most favored, additionally allowed, generously allowed, and disallowed regions.

Interestingly, both FUBP1 N-box-RRM interfaces show only limited interdigitation of the hydrophobic side chains, consistent with the modest binding affinity in the micromolar range.

In a recent survey of The Cancer Genome Atlas (TCGA), FUBP1 was noted for its particularly high rate of non-synonymous mutations in low-grade gliomas.[1] To learn about the mechanistic impact of such mutations, we systematically searched cancer mutation databases and identified 26 disease-related single-nucleotide variants (SNVs) within the FUBP1 N-box (Figure S4B). Five candidate mutations (A38D, A43E, K44R, I45F, and G47C) were selected by considering the magnitude of chemical shift changes occurring in the NMR titration of FUBP1$^{N\text{-}box}$ with U2AF2^{RRM2} (Figures S3B–S3D and S4B). In addition, we included L35V, which has been shown to weaken the FUBP1-PUF60 interaction.[61] NMR analysis revealed that A38D strongly impairs U2AF2 binding (Figures 3I and S4C–S4G). This is consistent with our structure in which A38 forms the core of the hydrophobic binding interface between FUBP1 N-box and U2AF2. A bulkier negatively charged side chain in this position is expected to introduce steric and electrostatic repulsion at the binding interface. Residue A38 in FUBP1 was also required for binding to PUF60 in a mutational study,[61] whereas L35V, which also affected the FUBP1-PUF60 interaction in that study, did not impair the interaction of FUBP1 with U2AF2 RRM2 (Figures S4C and S4D). A significant weakening of the U2AF2-FUBP1 interaction by A38D in the full-length context was also confirmed in cells using BRET (Figures 3C, 3J, and S2M). Here, some residual binding between FUBP1^{A38D} and U2AF2 was observed, probably because both proteins remain in proximity through binding to the same pre-mRNAs. As expected, A38D does not affect FUBP1-SF1 binding, which occurs via the A/B boxes (Figures 3C, S4H, and S4I). In summary, our experiments demonstrate that FUBP1 interacts directly with U2AF2 and SF1 via its N-terminal N-box and C-terminal A/B boxes, respectively. The former interaction is severely impaired by a cancer-associated mutation in FUBP1.

FUBP1 promotes U2AF2 binding to 3′ splice sites

To investigate the impact of FUBP1 on E complex formation, we monitored U2AF2 binding to RNA using *in vitro* iCLIP.[28] To this end, we designed a pool of short RNA transcripts (182 nt) representing ~2,000 natural 3′ splice sites from human transcripts, which we mixed with recombinant U2AF2^{RRM12} (see STAR Methods). Remarkably, addition of recombinant full-length FUBP1 (FUBP1^{FL}) results in stronger binding of U2AF2^{RRM12} to virtually all 3′ splice sites in the transcript pool (Figures 4A, 4B, and S5A–S5C; Table S1). The *in vivo* pattern of U2AF2 binding can thereby be reproduced *in vitro* in the presence of full-length FUBP1 (Figure 4C). The widespread effects are in contrast to those of our previous findings using *in vitro*-translated FUBP1, which affected only a few U2AF2 binding sites.[28] Hence, our updated experiments indicate that FUBP1 acts globally to stabilize U2AF2 binding. We find that this effect is dependent on FUBP1 concentration and is directly linked to the number of FUBP1 binding motifs upstream of the BP (Figure 4D). To confirm these findings in longer transcripts, we repeated the experiment with a pool of eight *in vitro* transcripts (2.0–5.7 kb; Figures S5D and S5E; Table S1). Indeed, addition of recombinant full-length FUBP1 increases the strength of U2AF2^{RRM12} binding at 3′ splice sites (Figures 4E and S5F) and thereby reproduces the *in vivo* binding pattern of U2AF2 (Figure 4F). Notably, this effect is considerably reduced with FUBP1$^{\Delta N}$ (impaired U2AF2 interaction), and it is completely abolished with FUBP1^{N74} (lacking KH domains). This highlights the importance of the N-box in FUBP1 for directly interacting with U2AF2 as well as of FUBP1's RNA binding for the stabilization of U2AF2 (Figures 4F and S5F). Together, this indicates that the interaction of FUBP1 with both pre-mRNA and U2AF2 globally promotes U2AF2 binding at the 3′ splice site during early spliceosomal assembly.

FUBP1 is critical for the splicing of long introns

To investigate the impact of FUBP1 on splicing, we generated a *FUBP1* knockout (KO) RPE1 cell line using CRISPR-Cas9 genome engineering (Figures 5A and S5G) and performed RNA-seq. *MYC* gene expression was unaltered, suggesting that it is not controlled by FUBP1 in RPE1 cells (Figure S5H). Next, we examined transcriptome-wide splicing and found 1,041 significant splicing changes, including 399 cassette exons (Figure 5B; Tables S1 and S3). Consistent with a role in splice site recognition, *FUBP1* KO preferentially leads to exon skipping (276 [69%] with delta percent spliced in [ΔPSI] < −0.1).

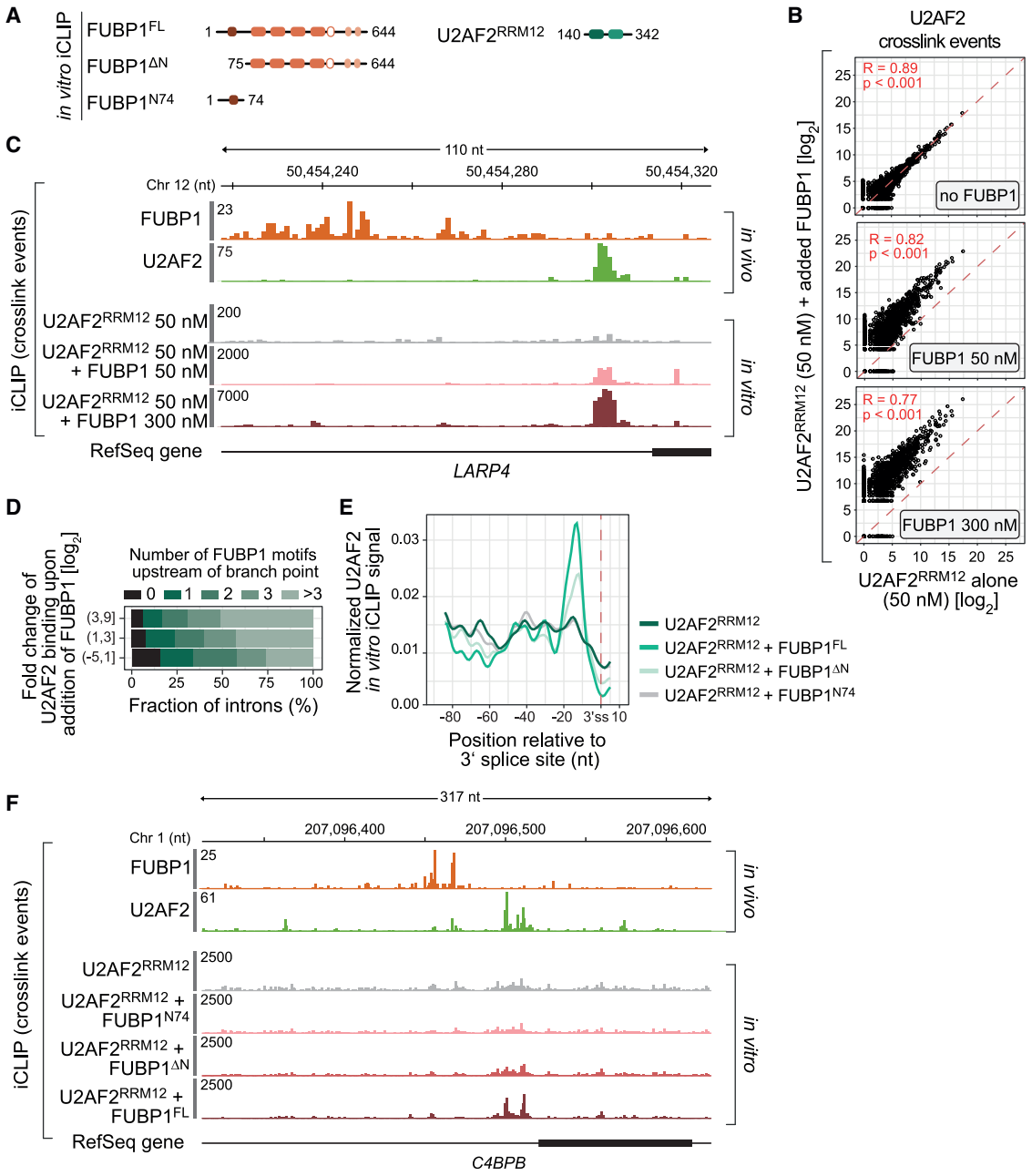


Figure 4. FUBP1 stabilizes U2AF2 binding at 3′ splice sites *in vitro*

(A) Overview of FUBP1 protein variants used in *in vitro* iCLIP experiments.

(B) Scatterplot of *in vitro* iCLIP signal in U2AF2 binding sites of $U2AF2^{RRM12}$ alone and upon addition of full-length FUBP1 on a pool of 1,998 *in vitro* transcripts.

(C) Genome browser view of *LARP4* mRNA displaying *in vivo* iCLIP for FUBP1 and U2AF2 and *in vitro* iCLIP on the respective *in vitro* transcript for U2AF2 alone and after addition of full-length FUBP1.

(D) Number of FUBP1 binding motifs upstream of the BP ([−100 nt; −26 nt]) in relation to the log_2-transformed fold change of $U2AF2^{RRM12}$ binding upon addition of full-length FUBP1 for 1,504 3′ splice sites in the *in vitro* transcripts.

(E) Metaprofile of U2AF2 binding at 3′ splice sites from *in vitro* iCLIP with long *in vitro* transcripts[28] and $U2AF2^{RRM12}$ alone and after addition of $FUBP1^{FL}$, $FUBP1^{N74}$, or $FUBP1^{\Delta N}$. iCLIP signals were normalized by spike-in and averaged per nucleotide over all introns (n = 21).

(F) Genome browser view of *C4BPB* mRNA displaying *in vivo* iCLIP for FUBP1 and U2AF2 and *in vitro* iCLIP for $U2AF2^{RRM12}$ alone and after addition of $FUBP1^{FL}$, $FUBP1^{N74}$ or $FUBP1^{\Delta N}$.

CellPress
OPEN ACCESS

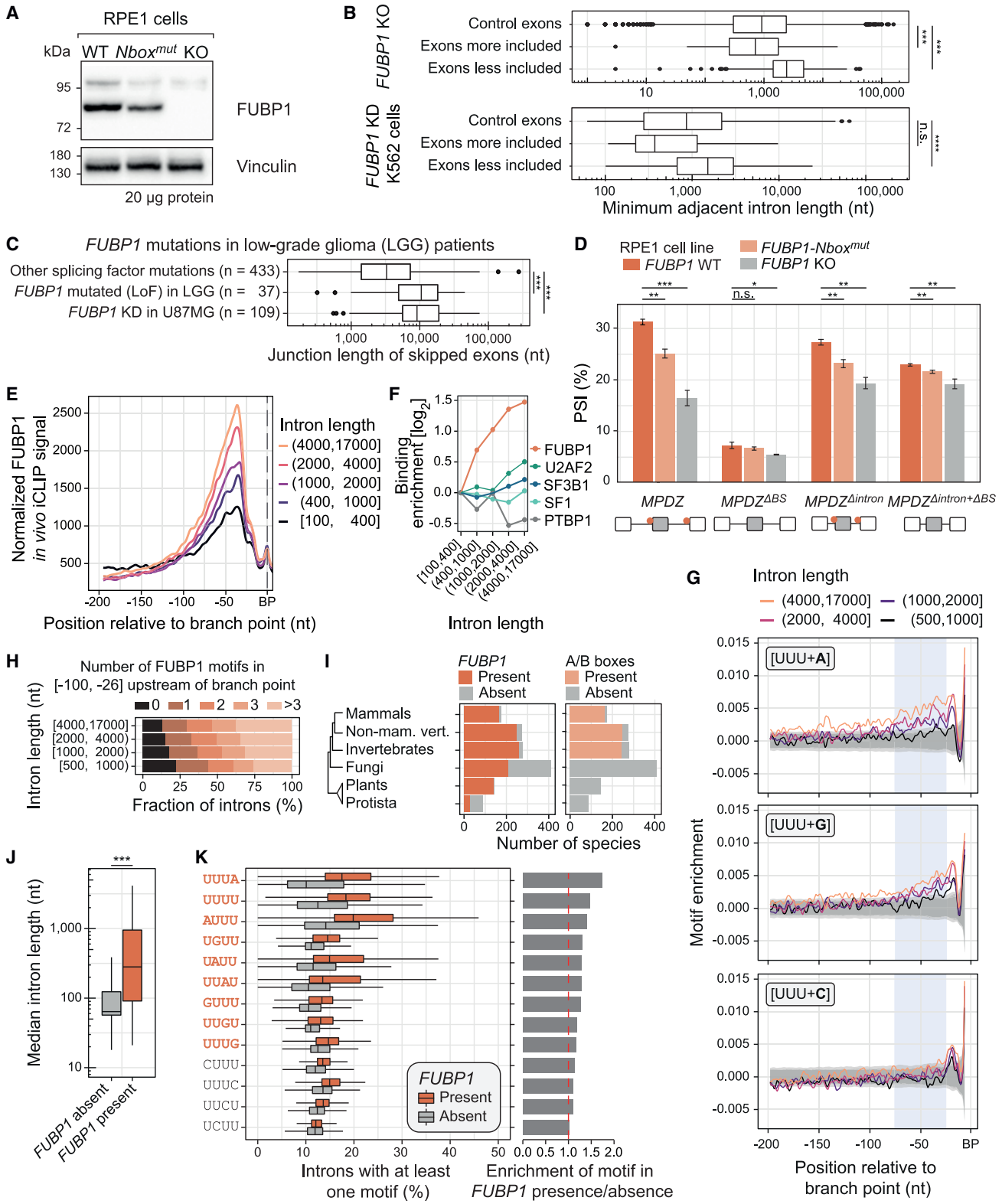


Figure 5. FUBP1 binds stronger to long introns and regulates exons flanked by long introns

(A) Western blot of FUBP1 in wild-type (WT), *FUBP1-Nbox*^{mut} mutant, and *FUBP1* KO RPE1 cells (Figure S5G). Vinculin acts as loading control.

(B) Minimum adjacent intron length for cassette exons more or less included upon *FUBP1* KO in RPE1 cells (n = 123/276) and *FUBP1* knockdown in K562 cells (n = 30/143) compared to unchanged control exons (RPE1, n = 10,301; K562, n = 1,910). ***p < 0.001, ****p < 0.0001, n.s., not significant.

(legend continued on next page)

A closer inspection revealed that the fate of an exon is related to the length of the flanking introns: decreased inclusion in *FUBP1* KO cells is typically observed for exons that are flanked by longer introns, compared with exons with increased or unchanged inclusion (Figure 5B, top). Most affected exons are alternative exons, but we observed the same effect for regulated constitutive exons (Figure S5I). Importantly, the effect on long introns can be recapitulated in ENCODE[62,63] data on *FUBP1* knockdown cells (Figure 5B, bottom). To test whether this depends on the interaction with U2AF2, we generated a *FUBP1-Nbox*^{mut} mutant with a targeted deletion of A38 and neighboring amino acids in the endogenous *FUBP1* gene in RPE1 cells (Figures 5A and S5G). Although overall fewer cassette exons are regulated in this mutant (n = 81), exons are predominantly skipped (n = 45), and these are flanked by longer introns (Figure S5J). Together, these data reveal that FUBP1 is important for the splicing of long introns and suggest a functional role for the N-box in this process.

To investigate whether *FUBP1* mutations in tumor cells affect splicing, we analyzed data from glioma patients.[1] Intriguingly, we found that skipped exons in patients with *FUBP1* loss-of-function mutations have longer adjacent introns than exons dysregulated in patients harboring other splicing factor mutations (Figures 5C and S6A). The effect is also evident upon *FUBP1* knockdown in the glioblastoma cell line U87MG from the same study (Figure 5C). Together, these data strongly suggest that FUBP1 plays a role in the efficient splicing of long introns, thereby affecting the inclusion of adjacent exons.

To validate the role of FUBP1 for long introns, we constructed a minigene for the alternative exon 18 in the *MPDZ* transcript, which is skipped upon *FUBP1* KO in RPE1 cells. The minigene comprises the alternative exon with the flanking constitutive exons and intervening long introns (>2.4 kb). *In vivo* iCLIP data show that FUBP1 binds at both 3′ splice sites, which was confirmed *in vitro* by EMSA with FUBP1$^{N\text{-box+KH}}$ (aa 1–457; Figures S6B and S6C). We observed a marked decrease of alternative exon inclusion from the *MPDZ* minigene in *FUBP1* KO (16% inclusion) and an intermediate effect (25%) in *FUBP1-Nbox*^{mut} cells, compared with wild-type (WT) cells (31% inclusion; Figures 5D, S6B, and S6D). Upon mutation of the FUBP1 binding sites, the exon showed reduced inclusion (7%) and did not change in the *FUBP1* KO. If the introns were shortened but the FUBP1 binding sites retained, the effect of *FUBP1* KO or mutation was reduced, albeit still present, consistent with the notion that the intron is still perceived as long due to the presence of FUBP1 binding site. By contrast, if the FUBP1 binding sites were also removed, exon inclusion no longer responded to *FUBP1* KO or *FUBP1-Nbox*^{mut}, highlighting that FUBP1 binding is specifically required for the long-intron variant.

Intriguingly, the changes at long introns are linked to FUBP1 binding. We found a substantial increase in FUBP1 binding at the 3′ splice sites of longer introns, both in absolute terms and relative to other splicing factors (Figures 5E and 5F). Differential FUBP1 binding was not observed for other exon-intron-related features, such as splice site, Py tract, and BP strength (Figures S6E–S6H). Furthermore, longer introns exhibit a marked enrichment of FUBP1 motifs upstream of the BP (Figures 5G and 5H). By contrast, random motif occurrences or splice site strength are independent of intron length (Figures S6I and S6J). Moreover, long introns were previously observed to preferentially locate to the nuclear periphery and exhibit a differential GC content architecture.[64,65] Indeed, we found that the occurrence of FUBP1 binding motifs correlates with the GC content architecture (Figures S6K–S6M). Furthermore, FUBP1 binds stronger to introns located in the nuclear periphery (Figure S6N) and to splice sites of exons with differential GC content architecture (Figures S7A–S7C). Further analysis indicated that both intron length and differential GC content architecture affect FUBP1 binding (Figure S7D).

Although splicing is an ancient molecular mechanism, gene architecture and especially intron length are subject to substantial evolutionary change (Figure S7E). We hypothesized that FUBP1 is present throughout Eukaryota and that lineage-specific losses or modifications of FUBP1 are accompanied by changes in average intron length. Indeed, we find overall that FUBP1 is well conserved. Although losses do occur, they are mostly observed in taxa with short introns such as protozoa and fungi (Figures 5I and 5J). Species with FUBP1 consistently harbor more FUBP1 motifs at their 3′ splice sites (Figure 5K). By contrast, U-rich motifs interspersed with C, which do not accumulate in the region of

(C) Junction length for less-included exons in RNA-seq from glioma patients with *FUBP1* loss-of-function (LoF) mutations, from a *FUBP1* siRNA knockdown in U87MG cells, and from *SF3B1*/*U2AF1*/*SRSF2* hotspot mutations and *RBM10* LoF mutation in different cancer patient samples. ***p < 0.001.

(D) Changes of exon inclusion (n = 3) in *FUBP1* WT, *FUBP1-Nbox*^{mut}, and *FUBP1* KO RPE1 cell lines upon intron shortening and/or removal of FUBP1 binding sites in the *MPDZ* minigene (Figure S6B). Data are represented as mean ± SD. Significance was determined by a two-sided Student's t-test with Benjamini-Hochberg correction. Red dots represent FUBP1 binding sites. *p < 0.05, **p < 0.01, ***p < 0.001, n.s., not significant.

(E) Metaprofile showing FUBP1 cross-link events relative to branch point for various intron lengths. iCLIP signals were normalized for expression and averaged per nucleotide over all introns.

(F) Quantification of binding signal based on area-under-the-curve (AUC) in main binding regions (see STAR Methods for details). Binding enrichment is defined as $\log_2$ fold change of AUC over AUC of introns with length in (100 nt, 400 nt).

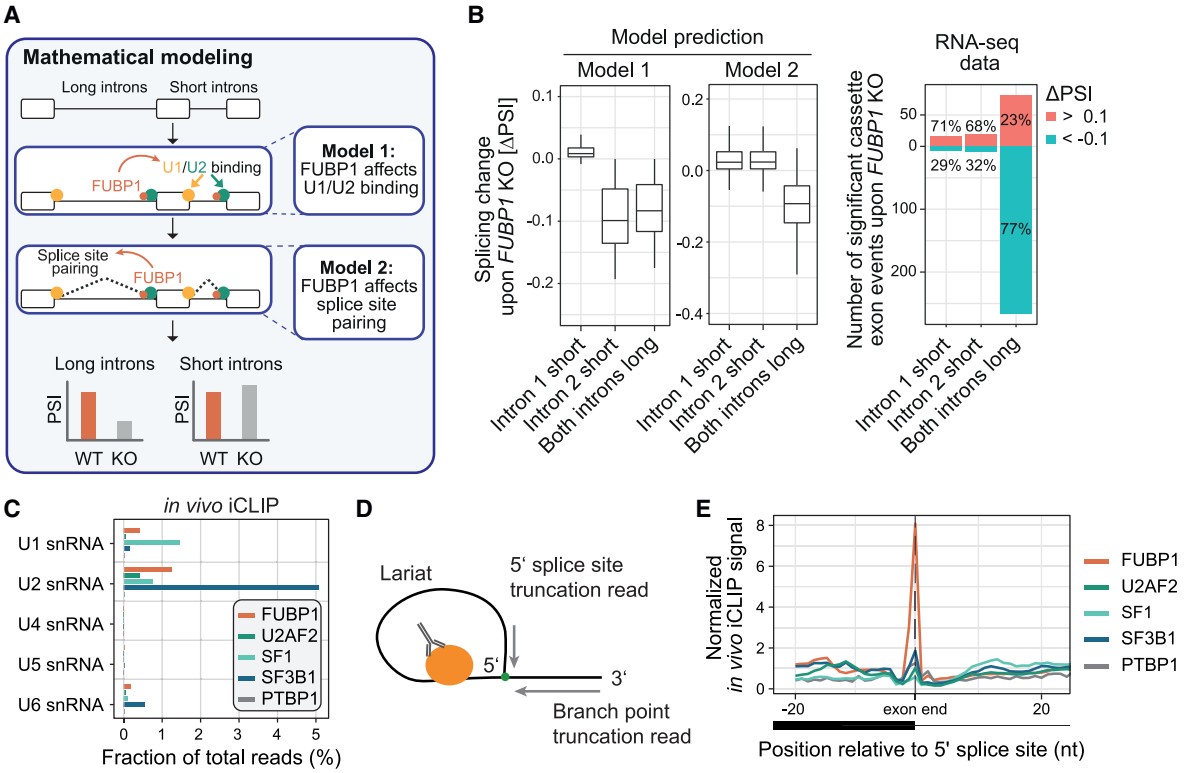
(G) Positional enrichment of FUBP1 binding motifs and control motifs relative to branch point and for various intron lengths. UUU+A/G/C, sets of four 4-mers containing UUU interspersed at any position with A/G/C. NNNN, 100,000 sets of random combinations of four 4-mers. 4-mer frequencies were calculated position-wise upstream of the BP and compared with average 4-mer frequencies in intronic control region.

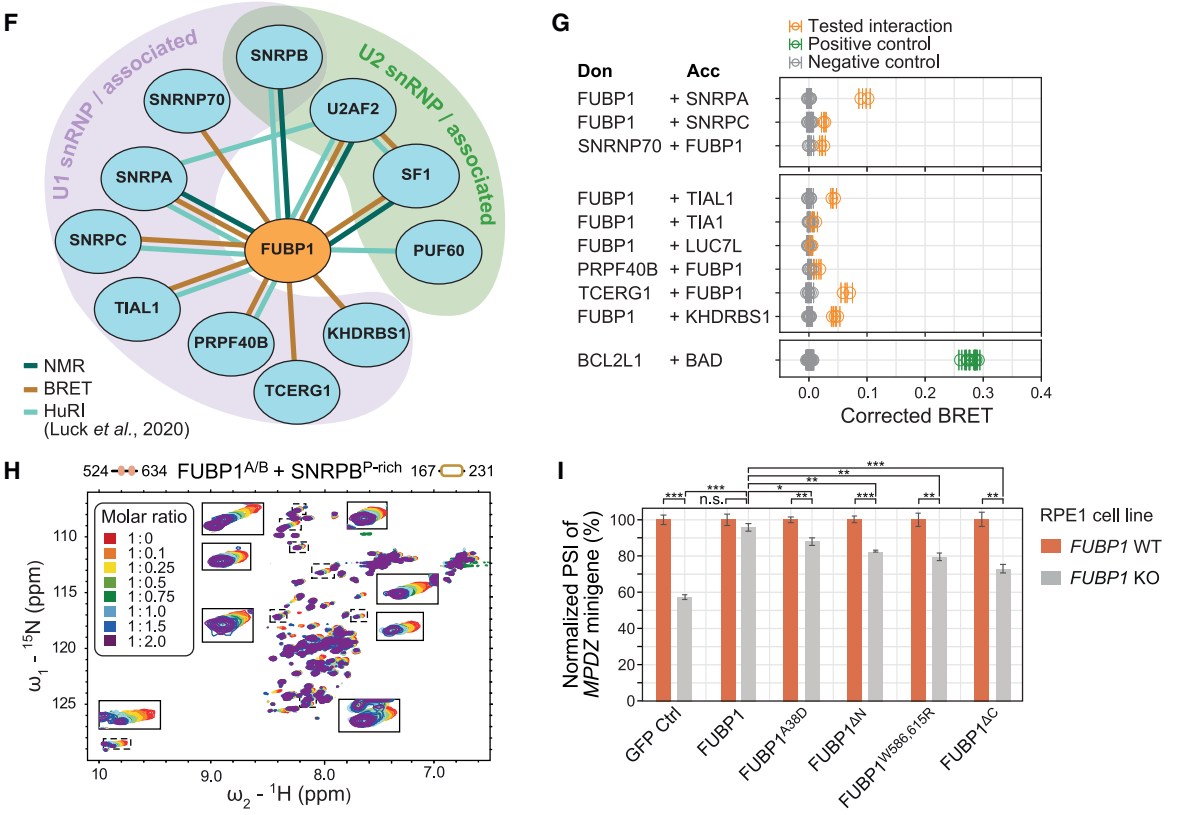
(H) Number of FUBP1 binding motifs upstream of the BP ([−100 nt; −26 nt]) for various intron lengths ([500, 1,000), n = 24,564 introns; [1,000, 2,000], n = 32,251 introns; [2,000, 4,000], n = 31,734 introns; [4,000, 17,000], n = 38,692).

(I) Phylogenetic profile of FUBP1. Tree indicates taxonomic range scanned for presence of FUBP1 orthologs. Fractions of species harboring ortholog to human FUBP1 (left) and carrying the A/B boxes (right) are shown.

(J) FUBP1 presence compared to median intron length per species. ***p < 0.001.

(K) Percentage of introns with at least one FUBP1 motif or control motifs present in 25-nt window located 25 nt upstream of the 3′ splice site.





(legend on next page)

FUBP1 binding (Figure 5G), are least enriched in species with FUBP1. Comparing FUBP1's domain architecture across eukaryotic evolution, we find that C-terminal A/B boxes are an animal-specific innovation. Their appearance in evolution is associated with an overall increase in intron length in animals compared with other eukaryotes (Figure 5I). Together, this suggests that FUBP1 binding to its RNA motifs and its protein-protein interfaces play important roles in the splicing of long introns.

FUBP1 interacts with both splice sites suggesting a function in cross-intron bridging

To decipher the molecular mechanism of FUBP1 action, we developed a kinetic model of cassette exon splicing using ordinary differential equations (Figures 6A and S7F; Table S4). In line with our previous work,[66] we considered a scenario for "exon definition" in which the U1 and U2 snRNPs recognize the 5′ and 3′ regions flanking an exon as functional units. The subsequent splice site pairing by U1/U2 snRNP interaction across the intron, that is, intron definition, triggers splicing catalysis, which results in either cassette exon inclusion, skipping, or intron retention in the model. We first simulated the loss of FUBP1 in a model in which FUBP1 solely acts on initial U1/U2 snRNP binding to exons (exon definition). However, our simulations argue against a pure exon definition effect, as the model cannot recapitulate the splicing changes that occur upon *FUBP1* KO (Figure 6B; model 1). According to our experimental data, exons flanked by two long introns are typically skipped upon *FUBP1* KO, whereas exons flanked by at least one short intron tend to show slightly increased inclusion. Surprisingly, the experimental data are more consistent with an alternative model in which FUBP1 enhances the pairing of splice sites across long (but not short) introns during intron definition. The model predicts reduced exon inclusion upon *FUBP1* KO specifically for exons flanked by two long introns, whereas exons flanked by one short intron moderately increase, irrespective of whether it is located upstream or downstream (Figure 6B; model 2). These results also hold true in a modified model, in which exons are not defined as functional units, and intron splicing solely requires U1 and U2 binding to flanking splice sites (Figure S7G, "intron definition model"). Taken together, the experimental observations are consistent with the kinetic model, which assumes that FUBP1 differentially affects long introns by promoting splice site pairing and the formation of catalytically active spliceosomes across long introns.

To test this prediction, we investigated the cross-linking of FUBP1 to snRNAs, indicative of its presence at different stages of splicing. First, FUBP1 showed substantial cross-linking to U2 snRNA, consistent with FUBP1 binding upstream of the BP where the U2 snRNP replaces SF1, indicating that FUBP1 is present during A-complex formation (Figure 6C). More importantly, FUBP1 also cross-links to U1 snRNA, which binds to the 5′ splice site, suggesting that FUBP1 is present during the bridging of the 3′ and 5′ splice sites, either during initial exon definition or also at later stages of intron definition. The latter is further supported by the cross-linking of FUBP1 to U6 snRNA, which replaces U1 snRNA at the 5′ splice site prior to lariat formation (Figure 6C). Hence, FUBP1 might be involved in intron bridging throughout the splicing cycle. We next searched our iCLIP datasets for evidence that FUBP1 is still bound in the spliceosomal C complex when the lariat has formed after the first splicing reaction. It has been shown that reads from the lariat truncate at the position where the 5′ splice site is covalently linked to the BP and is detected as a single-nucleotide-wide peak at the 5′ splice site (Figure 6D).[68,69] Indeed, we observed a strong peak in read truncations for FUBP1 at the 5′ splice site, whereas there was almost no signal for the other splicing factors tested (Figure 6E). This suggests that FUBP1 is present from the early stages of spliceosome assembly until at least the first catalytic step of the splicing reaction.

To further investigate whether FUBP1 is actively involved in splice-site bridging, we searched available binary protein-protein interaction data from yeast two-hybrid screens.[67] These data confirmed that FUBP1 binds to U2AF2 (Figure 6F). We also found evidence for FUBP1 interacting with several U1-associated proteins (SNRPA, SNRPC, TIAL1, and PRPF40B) as well

Figure 6. FUBP1 interacts with U1 snRNP components

(A) Kinetic model of FUBP1's effects on alternative splicing quantitatively describes steady-state abundance of splice products for a three-exon gene in control and *FUBP1* KO conditions. Two model variants were analyzed, in which FUBP1 affects the initial exon definition step near long introns (model 1), and the subsequent splicing reaction, promoting the excision of long introns (model 2). See STAR Methods for details.

(B) Simulated splicing changes upon *FUBP1* KO reflect transcriptome-wide RNA-seq data assuming that FUBP1 affects splicing catalysis (model 2). To reflect the heterogeneity of exons in the human transcriptome, kinetic parameters of the model were chosen at random, giving rise to an ensemble of 10,000 *in silico* exons. *FUBP1* KO was simulated for each *in silico* exon, assuming that FUBP1 either enhances exon definition (model 1) or the rate of splicing (model 2) for long (but not short) introns (see STAR Methods for details). In the data, significantly regulated cassette exons were classified based on flanking intron lengths (<400 nt = short, ≥400 nt = long).

(C) Fraction of total reads mapping to snRNAs using custom reference consisting of snRNAs (n = 10), tRNAs (n = 22), and rRNAs (n = 6).

(D) Schematic description of three-way junction of intron lariats. cDNAs can truncate not at the original protein-RNA interactions site but rather at the three-way junction. These cDNAs either start from the intron end and truncate at the BP or, alternatively, start downstream of the 5′ splice site and truncate at the first nucleotide of the intron.

(E) Metaprofiles showing cross-link events of FUBP1, U2AF2, SF3B1, SF1, and PTBP1 relative to the 5′ splice site. iCLIP signals were normalized for expression and averaged per nucleotide.

(F) Comprehensive interaction network of FUBP1 based on NMR, BRET, and published yeast two-hybrid data.[67]

(G) BRET measurements between FUBP1 and subunits of the U1 snRNP complex as well as U1 snRNP-associated proteins along with positive and negative control pairs. Biological replicates are shown. Error bars represent SD of technical triplicates.

(H) NMR titration of $FUBP1^{A/B}$ with $SNRPB^{P\text{-}rich}$ up to a molar ratio of 1:2. Significant chemical shift changes are highlighted by boxes.

(I) Percent-spliced-in (PSI) of *MPDZ* minigene upon transfection of WT and *FUBP1* KO RPE1 cells with different FUBP1 constructs. Data are represented as mean ± SD. Significance was determined by a two-sided Student's t-test with Benjamini-Hochberg correction. *p < 0.05, **p < 0.01, ***p < 0.001, n.s., not significant.

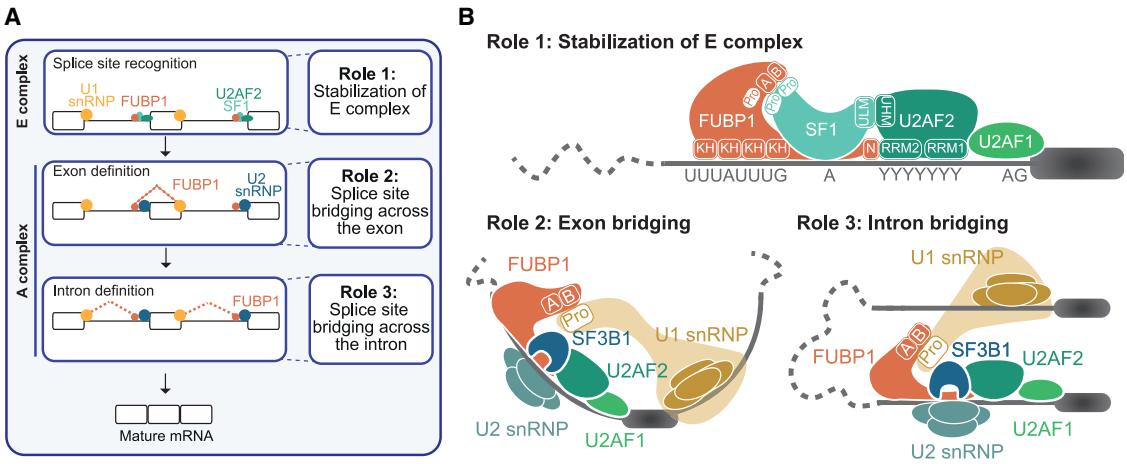


Figure 7. FUBP1 acts at multiple steps of early spliceosomal assembly

(A) The multiple roles of FUBP1 during spliceosomal complex assembly at the 3′ splice site.

(B) FUBP1 directly interacts with U2AF2, SF1, and additional U1/U2 snRNP components via distinct disordered interaction interfaces.

as with SNRPB, which is a member of the Sm protein ring in all snRNPs (Figure 6F). These and further interactions of FUBP1 with U1-associated proteins (TCERG1 and KHDRBS1) were confirmed using the BRET assay and/or NMR (Figures 6F, 6G, and S7H). Interestingly, several of the U1 snRNP-associated proteins harbor proline-rich regions, which potentially interact with the A/B boxes in FUBP1, similar to the FUBP1-SF1 interaction discussed above. Indeed, we observed significant changes in the NMR spectrum of FUBP1$^{A/B}$ upon the addition of a proline-rich peptide from SNRPB (Figure 6H), which were less pronounced with SNRPA and PRPF40B derivates (Figures S7I and S7J). This correlates well with the proline-rich region in SNRPB being much larger than in SNRPA or PRPF40B and thus avidity effects perhaps enhance the binding.

Finally, to confirm the importance of the FUBP1 A/B boxes and their role in splice-site bridging, we performed a complementation assay by expressing full-length GFP-FUBP1 and different mutants in both WT and *FUBP1* KO RPE1 cells. Effects on splicing were monitored using the co-transfected *MPDZ* minigene. As expected, GFP-FUBP1 complements the *FUBP1* KO cells and rescues *MPDZ* exon inclusion close to WT levels (Figures 6I, S7K, and S7L). Importantly, expression of GFP-FUBP1^{W586,615R} (mutations in the A/B boxes) or FUBP1$^{\Delta C}$ (complete deletion of the C terminus) impairs complementation in *FUBP1* KO cells. The same was also observed if the interaction with U2AF2 is perturbed by expressing either FUBP1^{A38D} (N-box mutation) or FUBP1$^{\Delta N}$ (complete deletion of the N terminus). Overall, these data demonstrate that both the A/B boxes and the N-box in FUBP1, which mediate the interactions with factors at the 5′ and 3′ splice sites, respectively, are functionally relevant for splicing.

DISCUSSION

FUBP1 is a general component of 3′ splice site definition

The recognition and pairing of splice sites, especially for the many long introns in the human transcriptome, are not well understood. In this study, we identified FUBP1 as a key component in 3′ splice site definition. We found that FUBP1 recognizes clustered U-rich elements interspersed by A or G that are present at virtually all 3′ splice sites and are most abundant for longer introns. Until now, four conserved intron-defining sequence motifs were known: the 5′ splice site motif, the BP sequence, the Py tract, and the 3′ splice site motif.[6] We propose the FUBP1 binding motif as a sequence signature that is relevant for spliceosomal assembly at long introns, which represent >80% of all human introns. Consistent with such a general role in splicing, FUBP1 has been detected in purified spliceosomes using mass spectrometry.[70–72]

We show that the four KH domains of FUBP1 recognize clustered arrays of binding motifs upstream of the BP. Multivalent interactions enhance binding affinity by avidity and enable the recognition of *cis*-elements in RNAs of variable length by combining individual KH-RNA motif interactions where multiple clustered RNA motifs may be separated by variable nucleotide linkers.[54] We find that the four KH domains are connected by flexible linkers, which facilitates scanning of extended RNA regions. The recognition of clustered RNA motifs by multidomain RBPs has been observed in IMP proteins and also involves four KH domains.[55] This suggests that KH domains working in concert might be a common mechanism for specifically recognizing clustered RNA motifs in extended RNA regions.

FUBP1 engages in multivalent interactions with 3′ and 5′ splice site components

We characterized two interfaces in FUBP1 that mediate protein-protein interactions: the N-box and the A/B boxes that are embedded in the intrinsically disordered N- and C-terminal regions of FUBP1, respectively. The N-box has been shown to interact with the RRM domain of PUF60 for regulation of transcription.[33,73,74] Here, we found that the FUBP1 N-box also binds to the RRM2 domain of U2AF2 and thereby mediates a functional interaction during pre-mRNA splicing. The N-box binds RRM2 opposite its RNA-binding surface, and thus, RNA

binding and FUBP1 binding do not compete. Notably, we have previously shown that the U2AF2 tandem domains adopt closed conformations and that RNA binding selects open arrangements.[15,29,75] Thus, binding of FUBP1 to the helical face of U2AF2 RRM2 might enhance RNA binding not only by stabilizing U2AF2 on the RNA but also by shifting the tandem RRM arrangements of U2AF2.

The A/B boxes of FUBP1 interact with intrinsically disordered proline-rich sequences within several U1 and U2 snRNP-associated proteins. This matches observations on the A/B boxes of the FUBP1 ortholog PSI in *Drosophila melanogaster*, which have been shown to bind to a proline-rich region in snRNP-U1-70K.[59] However, this region is not conserved in the human ortholog SNRNP70, and our BRET studies detected no such interaction between FUBP1 and SNRNP70. In general, linear motifs in proline-rich regions are recognized by structured regions such as WW or SH3 domains.[76] These interactions are generally weak but often enhanced by multivalent interactions.[77–81] Interestingly, the A/B boxes are unique to the FUBP family and appear to be unstructured regions in the ortholog PSI.[59] It will be interesting to learn how prevalent such an atypical mode of proline-rich sequence binding is and how it impacts cellular function.

FUBP1 contributes to spliceosome formation and guides the splicing of long introns

One important question is why FUBP1 is particularly relevant for long introns. Clearly, the splicing of long introns is difficult to achieve. For instance, it has been reported that exons flanking long introns are less included,[82,83] and that the splice sites of longer introns are stronger.[84,85] Consequently, longer introns require more complex regulation, such as the switch from initial exon definition to cross-intron spliceosomal complexes.[84,86] During exon definition, splice sites are recognized and paired across the exon, which is thereby defined as a functional unit. During the subsequent switch to intron definition, the complex shifts to a cross-intron pairing of splice sites (Figure 7). Our data suggest that FUBP1 acts at both steps. We propose that during exon definition, FUBP1 stabilizes U2AF2 and SF1 at the 3′ splice site. FUBP1 can thus strengthen the initial recognition of 3′ splice sites via its multivalent interactions with U2AF2, SF1, and pre-mRNA. The stabilization by FUBP1 and its interactions with the U1 snRNP across the exon might thus contribute to splice site recognition during exon definition.[86,87]

The interactions between FUBP1 and U1 snRNP components might also be relevant after the switch from exon definition to cross-intron pairing. Consistent with this model, we found that FUBP1 is still present at splice sites until the lariat is formed. In fact, FUBP1 forms cross-links to the U6 snRNA, which replaces U1 snRNA at the 5′ splice site. This indicates a role for FUBP1 in intron bridging during spliceosomal B-complex formation, particularly for long introns, as our experimental data and kinetic modeling suggest.

Several mechanisms and contributions to splice site bridging have been suggested, for example, the interactions between U1 and U2 snRNP proteins and RNA components[88–90] and the U2AF-associated RNA helicase UAP56.[91] It is conceivable that multiple contact sites act in concert to generate sufficient avidity to bring the splice sites together. Our data suggest that FUBP1—through multivalent interactions with pre-mRNA, proteins, and snRNAs located at the 5′ and 3′ splice sites—adds to these contacts throughout the splicing cycle. This is most pertinent for long introns harboring multiple FUBP1 *cis*-regulatory motifs.

In conclusion, we identify FUBP1 as a general splicing factor that ubiquitously binds at 3′ splice sites by means of a hitherto unknown *cis*-regulatory RNA sequence motif. The binding of FUBP1 and its interactions with multiple U1 and U2 snRNP components are pertinent to the efficient splicing of long introns.

Limitations of the study

Uridines are particularly prone to UV cross-linking, which can introduce bias to motif identification by iCLIP. However, we observed similar motifs using methods that do not involve UV cross-linking (NMR spectroscopy, ITC, and EMSA); therefore, we are confident that our conclusions in this regard are valid.

Upon depletion of FUBP1 in our KO or knockdown cell lines, other factors (such as the close paralog KHSRP) might, to some extent, take on the role of FUBP1. Together with cellular quality control mechanisms that degrade mis-spliced transcripts, this might reduce the effects of FUBP1 perturbation that we observed in our RNA-seq analysis. We might clarify such effects in the future by combing acute depletion of FUBP1 by means of degron tags with analysis of nascent RNA.

U2AF2 RRM2 and FUBP1 N-box interact with weak affinity in the micromolar range. Although it is likely that the simultaneous binding of U2AF2 and FUBP1 to the RNA further stabilizes this interaction, we cannot exclude the involvement of other factors.

In general, introns may be characterized by a multitude of features, among which length is just one. For example, intron length is known to correlate with elevated differential GC content and overall lower intron and exon GC content.[65] In addition, genes with longer introns have been shown to preferentially localize to the nuclear periphery,[64] and their transcripts therefore might interact with different splicing factors than for genes at the nuclear center. The question of whether these attributes rather complement each other or are causally related remains to be answered.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - RPE1 cell lines and culture conditions
  - HeLa cell line and culture conditions
  - HEK cell line and culture conditions
  - Recombinant protein expression
- METHOD DETAILS
  - Establishing *FUBP1* KO/*Nbox*^{mut} cell lines

- Immunoblotting
- RPE1 RNA-seq
- HeLa RNA-seq
- Semi-quantitative RT-PCR
- *In vivo* iCLIP
- *In vitro* iCLIP
- Protein expression and purification
- NMR spectroscopy
- *In vitro* binding assays
- BRET

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Preprocessing of RNA-seq data
  - Preprocessing of *in vivo* iCLIP data
  - Metaprofiles for *in vivo* iCLIP data
  - iCLIP binding site definition (peak calling)
  - Saturation analysis
  - Motif enrichment for *in vivo* iCLIP
  - Motif enrichment upstream of branch points
  - Abundance of FUBP1 motif at 3′ splice sites
  - Analysis of *in vitro* iCLIP data
  - Intron length analyses of RNA-seq data
  - ENCODE data analysis
  - Splicing changes upon FUBP1 LoF mutations
  - Mutations in FUBP1 in cancer patients
  - Scoring of splice site features
  - Evolutionary analyses
  - Analysis of RBP crosslinking to snRNAs
  - Subnuclear distribution of FUBP1-bound genes
  - Mathematical modeling

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.molcel.2023.07.002.

ACKNOWLEDGMENTS

We thank all the members of the Luck, Sattler, and König labs for their help and discussion. We thank Malgorzata Rogalska and Juan Valcárcel for discussions and comments on the manuscript, Philipp Trepte and the Wanker group for sharing protocols and reagents and for help in setting up BRET assays, Christian Schäfer for help with BRET assays, Eric Schumbera for help with BRET data processing, Fridolin Kielisch for help with statistical analyses, Mario Keller for bioinformatics advice, André Mourão for SNRPB$^{P\text{-}rich}$ plasmid, Sam Asami and Gerd Gemmecker for support with NMR experiments, Manuel Kaulich for reagents, and Chris Smith and Jernej Ule for PTBP1-RB40 antibody and resequencing. We thank Adrian Neal for editing and commenting on the manuscript. We thank the Core Facilities at IMB, in particular Protein Production, Microscopy, Bioinformatics, Genomics, and Flow Cytometry.

We acknowledge IMB Genomics Core Facility and its NextSeq 500 sequencer (funded by the Deutsche Forschungsgemeinschaft [DFG, German Research Foundation] INST 247/870-1 FUGG) and access to NMR spectrometers at Bavarian NMR Center. This work was supported by DFG grants to K.L. (LU 2568/1-1; SFB1551 Project no. 464588647), J.K. (SPP1935 Project no. 273941853, KO4566/2-1, SFB1551 Project No. 464588647, TRR 319 Project no. 439669440, and GRK2526/1 Project no. 407023052), K.Z. (SPP1935 Project no. 273941853), S.L. (LE 3473/2–3), and M.S. (SPP1935 Project no. 273941853, SA823/10-1, and SFB1035 Project no. 201302640). C.H. acknowledges the Fonds der Chemischen Industrie for Kekulé fellowship, and S.M.-L. acknowledges EU Horizon 2020 Research and Innovation program under the Marie Skłodovska-Curie grant agreement No. 792692. J.S. acknowledges a PhD stipend from IMB's collaborative research initiative.

AUTHOR CONTRIBUTIONS

S.E., C.B., A. Busch, and A.D.L. performed the bioinformatic analyses. C.H., H.-S.K., and S.M.-L. performed the structural, biophysical, and biochemical experiments and analyses. M.M. Mulorz, A. Buchbender, F.X.R.S., L.L.A., H.H., K.T., and M.M. Möckel performed the functional genomics, *in vitro* iCLIP, and minigene reporter experiments. D.H., J.S., and M.W. performed the BRET experiments. P.K. and S.L. performed the mathematical modeling. I.E. performed the evolutionary analysis. S.E., C.H., M.M. Mulorz, K.Z., K.L., M.S., and J.K. designed the study and wrote the manuscript. All authors read and commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 4, 2023
Revised: May 19, 2023
Accepted: July 3, 2023
Published: July 27, 2023

REFERENCES

1. Seiler, M., Peng, S., Agrawal, A.A., Palacino, J., Teng, T., Zhu, P., Smith, P.G., Cancer; Genome; Atlas; Research Network, Buonamici, S., and Yu, L. (2018). Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. Cell Rep. *23*, 282–296.e4. https://doi.org/10.1016/j.celrep.2018.01.088.
2. Bonnal, S.C., López-Oreja, I., and Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer – implications for care. Nat. Rev. Clin. Oncol. *17*, 457–474. https://doi.org/10.1038/s41571-020-0350-x.
3. Gebauer, F., Schwarzl, T., Valcárcel, J., and Hentze, M.W. (2021). RNA-binding proteins in human genetic disease. Nat. Rev. Genet. *22*, 185–198. https://doi.org/10.1038/s41576-020-00302-y.
4. Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nat. Rev. Mol. Cell Biol. *18*, 655–670. https://doi.org/10.1038/nrm.2017.86.
5. Wilkinson, M.E., Charenton, C., and Nagai, K. (2020). RNA splicing by the spliceosome. Annu. Rev. Biochem. *89*, 359–388. https://doi.org/10.1146/annurev-biochem-091719-064225.
6. Wahl, M.C., Will, C.L., and Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell *136*, 701–718. https://doi.org/10.1016/j.cell.2009.02.009.
7. Papasaikas, P., and Valcárcel, J. (2016). The spliceosome: the ultimate RNA chaperone and sculptor. Trends Biochem. Sci. *41*, 33–45. https://doi.org/10.1016/j.tibs.2015.11.003.
8. Berglund, J.A., Abovich, N., and Rosbash, M. (1998). A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. Genes Dev. *12*, 858–867. https://doi.org/10.1101/gad.12.6.858.
9. Liu, Z., Luyten, I., Bottomley, M.J., Messias, A.C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Krämer, A., and Sattler, M. (2001). Structural basis for recognition of the intron branch site RNA by splicing factor 1. Science *294*, 1098–1102. https://doi.org/10.1126/science.1064719.
10. Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Krämer, A., and Sattler, M. (2003). Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. Mol. Cell *11*, 965–976. https://doi.org/10.1016/s1097-2765(03)00115-1.
11. Kielkopf, C.L., Rodionova, N.A., Green, M.R., and Burley, S.K. (2001). A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. Cell *106*, 595–605. https://doi.org/10.1016/s0092-8674(01)00480-9.
12. Wu, S., Romfo, C.M., Nilsen, T.W., and Green, M.R. (1999). Functional recognition of the 3′ splice site AG by the splicing factor U2AF35. Nature *402*, 832–835. https://doi.org/10.1038/45590.

13. Merendino, L., Guth, S., Bilbao, D., Martínez, C., and Valcárcel, J. (1999). Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3′ splice site AG. Nature *402*, 838–841. https://doi.org/10.1038/45602.

14. Agrawal, A.A., Salsi, E., Chatrikhi, R., Henderson, S., Jenkins, J.L., Green, M.R., Ermolenko, D.N., and Kielkopf, C.L. (2016). An extended U2AF(65)–RNA-binding domain recognizes the 3′ splice site signal. Nat. Commun. *7*, 10950. https://doi.org/10.1038/ncomms10950.

15. Mackereth, C.D., Madl, T., Bonnal, S., Simon, B., Zanier, K., Gasch, A., Rybin, V., Valcárcel, J., and Sattler, M. (2011). Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. Nature *475*, 408–411. https://doi.org/10.1038/nature10171.

16. Zamore, P.D., and Green, M.R. (1989). Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. Proc. Natl. Acad. Sci. USA *86*, 9243–9247. https://doi.org/10.1073/pnas.86.23.9243.

17. Berglund, J.A., Chua, K., Abovich, N., Reed, R., and Rosbash, M. (1997). The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. Cell *89*, 781–787. https://doi.org/10.1016/s0092-8674(00)80261-5.

18. Crisci, A., Raleff, F., Bagdiul, I., Raabe, M., Urlaub, H., Rain, J.-C., and Krämer, A. (2015). Mammalian splicing factor SF1 interacts with SURP domains of U2 snRNP-associated proteins. Nucleic Acids Res. *43*, 10456–10473. https://doi.org/10.1093/nar/gkv952.

19. Wahl, M.C., and Lührmann, R. (2015). SnapShot: spliceosome dynamics I. Cell *161*, 1474–1474e1. https://doi.org/10.1016/j.cell.2015.05.050.

20. Tholen, J., and Galej, W.P. (2022). Structural studies of the spliceosome: bridging the gaps. Curr. Opin. Struct. Biol. *77*, 102461. https://doi.org/10.1016/j.sbi.2022.102461.

21. Ule, J., and Blencowe, B.J. (2019). Alternative splicing regulatory networks: functions, mechanisms, and evolution. Mol. Cell *76*, 329–345. https://doi.org/10.1016/j.molcel.2019.09.017.

22. Zuo, P., and Maniatis, T. (1996). The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. Genes Dev. *10*, 1356–1368. https://doi.org/10.1101/gad.10.11.1356.

23. Saulière, J., Sureau, A., Expert-Bezançon, A., and Marie, J. (2006). The polypyrimidine tract binding protein (PTB) represses splicing of exon 6B from the beta-tropomyosin pre-mRNA by directly interfering with the binding of the U2AF65 subunit. Mol. Cell. Biol. *26*, 8755–8769. https://doi.org/10.1128/MCB.00893-06.

24. Soares, L.M.M., Zanier, K., Mackereth, C., Sattler, M., and Valcárcel, J. (2006). Intron removal requires proofreading of U2AF/3′ splice site recognition by DEK. Science *312*, 1961–1965. https://doi.org/10.1126/science.1128659.

25. Warf, M.B., Diegel, J.V., von Hippel, P.H., and Berglund, J.A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proc. Natl. Acad. Sci. USA *106*, 9203–9208. https://doi.org/10.1073/pnas.0900342106.

26. Tavanez, J.P., Madl, T., Kooshapur, H., Sattler, M., and Valcárcel, J. (2012). hnRNP A1 proofreads 3′ splice site recognition by U2AF. Mol. Cell *45*, 314–329. https://doi.org/10.1016/j.molcel.2011.11.033.

27. Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. Cell *152*, 453–466. https://doi.org/10.1016/j.cell.2012.12.023.

28. Sutandy, F.X.R., Ebersberger, S., Huang, L., Busch, A., Bach, M., Kang, H.-S., Fallmann, J., Maticzka, D., Backofen, R., Stadler, P.F., et al. (2018). In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. Genome Res. *28*, 699–713. https://doi.org/10.1101/gr.229757.117.

29. Voith von Voithenberg, L., Sánchez-Rico, C., Kang, H.-S., Madl, T., Zanier, K., Barth, A., Warner, L.R., Sattler, M., and Lamb, D.C. (2016). Recognition of the 3′ splice site RNA by the U2AF heterodimer involves a dynamic population shift. Proc. Natl. Acad. Sci. USA *113*. E7169–E7175. https://doi.org/10.1073/pnas.1605873113.

30. Kang, H.-S., Sánchez-Rico, C., Ebersberger, S., Sutandy, F.X.R., Busch, A., Welte, T., Stehle, R., Hipp, C., Schulz, L., Buchbender, A., et al. (2020). An autoinhibitory intramolecular interaction proof-reads RNA recognition by the essential splicing factor U2AF2. Proc. Natl. Acad. Sci. USA *117*, 7140–7149. https://doi.org/10.1073/pnas.1913483117.

31. Debaize, L., and Troadec, M.-B. (2019). The master regulator FUBP1: its emerging role in normal cell function and malignant development. Cell. Mol. Life Sci. *76*, 259–281. https://doi.org/10.1007/s00018-018-2933-6.

32. Duncan, R., Bazar, L., Michelotti, G., Tomonaga, T., Krutzsch, H., Avigan, M., and Levens, D. (1994). A sequence-specific, single-strand binding protein activates the far upstream element of c-myc and defines a new DNA-binding motif. Genes Dev. *8*, 465–480. https://doi.org/10.1101/gad.8.4.465.

33. Liu, J., Kouzine, F., Nie, Z., Chung, H.-J., Elisha-Feil, Z., Weber, A., Zhao, K., and Levens, D. (2006). The FUSE/FBP/FIR/TFIIH system is a molecular machine programming a pulse of *c-myc* expression. EMBO J. *25*, 2119–2130. https://doi.org/10.1038/sj.emboj.7601101.

34. Cukier, C.D., Hollingworth, D., Martin, S.R., Kelly, G., Díaz-Moreno, I., and Ramos, A. (2010). Molecular basis of FIR-mediated *c-myc* transcriptional control. Nat. Struct. Mol. Biol. *17*, 1058–1064. https://doi.org/10.1038/nsmb.1883.

35. Li, H., Wang, Z., Zhou, X., Cheng, Y., Xie, Z., Manley, J.L., and Feng, Y. (2013). Far upstream element-binding protein 1 and RNA secondary structure both mediate second-step splicing repression. Proc. Natl. Acad. Sci. USA *110*. E2687–E2695. https://doi.org/10.1073/pnas.1310607110.

36. Hwang, I., Cao, D., Na, Y., Kim, D.-Y., Zhang, T., Yao, J., Oh, H., Hu, J., Zheng, H., Yao, Y., and Paik, J. (2018). Far upstream element-binding protein 1 regulates LSD1 alternative splicing to promote terminal differentiation of neural progenitors. Stem Cell Reports *10*, 1208–1221. https://doi.org/10.1016/j.stemcr.2018.02.013.

37. Jacob, A.G., Singh, R.K., Mohammad, F., Bebee, T.W., and Chandler, D.S. (2014). The splicing factor FUBP1 is required for the efficient splicing of oncogene MDM2 pre-mRNA. J. Biol. Chem. *289*, 17350–17364. https://doi.org/10.1074/jbc.M114.554717.

38. Miro, J., Laaref, A.M., Rofidal, V., Lagrafeuille, R., Hem, S., Thorel, D., Méchin, D., Mamchaoui, K., Mouly, V., Claustres, M., and Tuffery-Giraud, S. (2015). FUBP1: a new protagonist in splicing regulation of the DMD gene. Nucleic Acids Res. *43*, 2378–2389. https://doi.org/10.1093/nar/gkv086.

39. Ni, X., Knapp, S., and Chaikuad, A. (2020). Comparative structural analyses and nucleotide-binding characterization of the four KH domains of FUBP1. Sci. Rep. *10*, 13459. https://doi.org/10.1038/s41598-020-69832-z.

40. Wang, H., Zhang, R., Li, E., Yan, R., Ma, B., and Ma, Q. (2022). Pan-cancer transcriptome and immune infiltration analyses reveal the oncogenic role of far upstream element-binding protein 1 (FUBP1). Front. Mol. Biosci. *9*, 794715. https://doi.org/10.3389/fmolb.2022.794715.

41. Elman, J.S., Ni, T.K., Mengwasser, K.E., Jin, D., Wronski, A., Elledge, S.J., and Kuperwasser, C. (2019). Identification of FUBP1 as a long tail cancer driver and widespread regulator of tumor suppressor and oncogene alternative splicing. Cell Rep. *28*, 3435–3449.e5. https://doi.org/10.1016/j.celrep.2019.08.060.

42. Wang, J., Schultz, P.G., and Johnson, K.A. (2018). Mechanistic studies of a small-molecule modulator of SMN2 splicing. Proc. Natl. Acad. Sci. USA *115*. E4604–E4612. https://doi.org/10.1073/pnas.1800260115.

43. König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of

hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17*, 909–915. https://doi.org/10.1038/nsmb.1838.

44. Buchbender, A., Mutter, H., Sutandy, F.X.R., Körtel, N., Hänel, H., Busch, A., Ebersberger, S., and König, J. (2020). Improved library preparation with the new iCLIP2 protocol. Methods *178*, 33–48. https://doi.org/10.1016/j.ymeth.2019.10.003.

45. Valcárcel, J., Gaur, R.K., Singh, R., and Green, M.R. (1996). Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected]. Science *273*, 1706–1709. https://doi.org/10.1126/science.273.5282.1706.

46. Singh, R., Valcárcel, J., and Green, M.R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science *268*, 1173–1176. https://doi.org/10.1126/science.7761834.

47. Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., and Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biol. *13*, R67. https://doi.org/10.1186/gb-2012-13-8-r67.

48. Gozani, O., Potashkin, J., and Reed, R. (1998). A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. Mol. Cell. Biol. *18*, 4752–4760. https://doi.org/10.1128/MCB.18.8.4752.

49. Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D.L., Sun, H., et al. (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol. Cell *36*, 996–1006. https://doi.org/10.1016/j.molcel.2009.12.003.

50. Llorian, M., Schwartz, S., Clark, T.A., Hollander, D., Tan, L.-Y., Spellman, R., Gordon, A., Schweitzer, A.C., de la Grange, P., Ast, G., and Smith, C.W.J. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat. Struct. Mol. Biol. *17*, 1114–1123. https://doi.org/10.1038/nsmb.1881.

51. Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., Zhou, J., Qiu, J., Jiang, L., Li, H., et al. (2014). Mechanisms for U2AF to define 3′ splice sites and regulate alternative splicing in the human genome. Nat. Struct. Mol. Biol. *21*, 997–1005. https://doi.org/10.1038/nsmb.2906.

52. Valverde, R., Edwards, L., and Regan, L. (2008). Structure and function of KH domains. FEBS J. *275*, 2712–2726. https://doi.org/10.1111/j.1742-4658.2008.06411.x.

53. Fukumura, K., Yoshimoto, R., Sperotto, L., Kang, H.-S., Hirose, T., Inoue, K., Sattler, M., and Mayeda, A. (2021). SPF45/RBM17-dependent, but not U2AF-dependent, splicing in a distinct subset of human short introns. Nat. Commun. *12*, 4910. https://doi.org/10.1038/s41467-021-24879-y.

54. Mackereth, C.D., and Sattler, M. (2012). Dynamics in multi-domain protein recognition of RNA. Curr. Opin. Struct. Biol. *22*, 287–296. https://doi.org/10.1016/j.sbi.2012.03.013.

55. Schneider, T., Hung, L.-H., Aziz, M., Wilmen, A., Thaum, S., Wagner, J., Janowski, R., Müller, S., Schreiner, S., Friedhoff, P., et al. (2019). Combinatorial recognition of clustered RNA elements by the multidomain RNA-binding protein IMP3. Nat. Commun. *10*, 2266. https://doi.org/10.1038/s41467-019-09769-8.

56. Siomi, H., Matunis, M.J., Michael, W.M., and Dreyfuss, G. (1993). The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. Nucleic Acids Res. *21*, 1193–1198. https://doi.org/10.1093/nar/21.5.1193.

57. Beuth, B., García-Mayoral, M.F., Taylor, I.A., and Ramos, A. (2007). Scaffold-independent analysis of RNA-protein interactions: the Nova-1 KH3-RNA complex. J. Am. Chem. Soc. *129*, 10205–10210. https://doi.org/10.1021/ja072365q.

58. Trepte, P., Kruse, S., Kostova, S., Hoffmann, S., Buntru, A., Tempelmeier, A., Secker, C., Diez, L., Schulz, A., Klockmeier, K., et al. (2018). LuTHy: a double-readout bioluminescence-based two-hybrid technology for quantitative mapping of protein-protein interactions in mammalian cells. Mol. Syst. Biol. *14*, e8071. https://doi.org/10.15252/msb.20178071.

59. Ignjatovic, T., Yang, J.-C., Butler, J., Neuhaus, D., and Nagai, K. (2005). Structural basis of the interaction between P-element somatic inhibitor and U1-70k essential for the alternative splicing of P-element transposase. J. Mol. Biol. *351*, 52–65. https://doi.org/10.1016/j.jmb.2005.04.077.

60. Labourier, E., Adams, M.D., and Rio, D.C. (2001). Modulation of P-element pre-mRNA splicing by a direct interaction between PSI and U1 snRNP 70K protein. Mol. Cell *8*, 363–373. https://doi.org/10.1016/s1097-2765(01)00311-2.

61. Chung, H.-J., Liu, J., Dundr, M., Nie, Z., Sanford, S., and Levens, D. (2006). FBPs are calibrated molecular tools to adjust gene expression. Mol. Cell. Biol. *26*, 6584–6597. https://doi.org/10.1128/MCB.00754-06.

62. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

63. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. *48*. D882–D889. https://doi.org/10.1093/nar/gkz1062.

64. Tammer, L., Hameiri, O., Keydar, I., Roy, V.R., Ashkenazy-Titelman, A., Custódio, N., Sason, I., Shayevitch, R., Rodríguez-Vaello, V., Rino, J., et al. (2022). Gene architecture directs splicing outcome in separate nuclear spatial regions. Mol. Cell *82*, 1021–1034.e8. https://doi.org/10.1016/j.molcel.2022.02.001.

65. Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. *1*, 543–556. https://doi.org/10.1016/j.celrep.2012.03.013.

66. Enculescu, M., Braun, S., Thonta Setty, S., Busch, A., Zarnack, K., König, J., and Legewie, S. (2020). Exon definition facilitates reliable control of alternative splicing in the RON proto-oncogene. Biophys. J. *118*, 2027–2041. https://doi.org/10.1016/j.bpj.2020.02.022.

67. Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408. https://doi.org/10.1038/s41586-020-2188-x.

68. Briese, M., Haberman, N., Sibley, C.R., Faraway, R., Elser, A.S., Chakrabarti, A.M., Wang, Z., König, J., Perera, D., Wickramasinghe, V.O., et al. (2019). A systems view of spliceosomal assembly and branchpoints with iCLIP. Nat. Struct. Mol. Biol. *26*, 930–940. https://doi.org/10.1038/s41594-019-0300-4.

69. Cordiner, R.A., Dou, Y., Thomsen, R., Bugai, A., Granneman, S., and Heick Jensen, T. (2023). Temporal-iCLIP captures co-transcriptional RNA-protein interactions. Nat. Commun. *14*, 696. https://doi.org/10.1038/s41467-023-36345-y.

70. Rappsilber, J., Ryder, U., Lamond, A.I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. Genome Res. *12*, 1231–1245. https://doi.org/10.1101/gr.473902.

71. Makarov, E.M., Owen, N., Bottrill, A., and Makarova, O.V. (2012). Functional mammalian spliceosomal complex E contains SMN complex proteins in addition to U1 and U2 snRNPs. Nucleic Acids Res. *40*, 2639–2652. https://doi.org/10.1093/nar/gkr1056.

72. Sharma, S., Kohlstaedt, L.A., Damianov, A., Rio, D.C., and Black, D.L. (2008). Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. Nat. Struct. Mol. Biol. *15*, 183–191. https://doi.org/10.1038/nsmb.1375.

73. Hsiao, H.-H., Nath, A., Lin, C.-Y., Folta-Stogniew, E.J., Rhoades, E., and Braddock, D.T. (2010). Quantitative characterization of the interactions among c-myc transcriptional regulators FUSE, FBP, and FIR. Biochemistry *49*, 4620–4634. https://doi.org/10.1021/bi9021445.

74. Liu, J., He, L., Collins, I., Ge, H., Libutti, D., Li, J., Egly, J.M., and Levens, D. (2000). The FBP interacting repressor targets TFIIH to inhibit activated transcription. Mol. Cell *5*, 331–341. https://doi.org/10.1016/s1097-2765(00)80428-1.

75. Huang, J.-R., Warner, L.R., Sanchez, C., Gabel, F., Madl, T., Mackereth, C.D., Sattler, M., and Blackledge, M. (2014). Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. J. Am. Chem. Soc. *136*, 7068–7076. https://doi.org/10.1021/ja502030n.

76. Macias, M.J., Wiesner, S., and Sudol, M. (2002). WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. FEBS Lett. *513*, 30–37. https://doi.org/10.1016/s0014-5793(01)03290-2.

77. Ball, L.J., Kühne, R., Schneider-Mergener, J., and Oschkinat, H. (2005). Recognition of proline-rich motifs by protein-protein-interaction domains. Angew. Chem. Int. Ed. Engl. *44*, 2852–2869. https://doi.org/10.1002/anie.200400618.

78. Zarrinpar, A., Bhattacharyya, R.P., and Lim, W.A. (2003). The structure and function of proline recognition domains. Sci. STKE *2003*, RE8. https://doi.org/10.1126/stke.2003.179.re8.

79. Kofler, M.M., and Freund, C. (2006). The GYF domain. FEBS J. *273*, 245–256. https://doi.org/10.1111/j.1742-4658.2005.05078.x.

80. Sudol, M. (1996). Structure and function of the WW domain. Prog. Biophys. Mol. Biol. *65*, 113–132. https://doi.org/10.1016/s0079-6107(96)00008-9.

81. Mayer, B.J. (2001). SH3 domains: complexity in moderation. J. Cell Sci. *114*, 1253–1263. https://doi.org/10.1242/jcs.114.7.1253.

82. Bell, M.V., Cowper, A.E., Lefranc, M.P., Bell, J.I., and Screaton, G.R. (1998). Influence of intron length on alternative splicing of CD44. Mol. Cell. Biol. *18*, 5930–5941. https://doi.org/10.1128/MCB.18.10.5930.

83. Fox-Walsh, K.L., Dou, Y., Lam, B.J., Hung, S.-P., Baldi, P.F., and Hertel, K.J. (2005). The architecture of pre-mRNAs affects mechanisms of splice-site pairing. Proc. Natl. Acad. Sci. USA *102*, 16176–16181. https://doi.org/10.1073/pnas.0508489102.

84. Dewey, C.N., Rogozin, I.B., and Koonin, E.V. (2006). Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. BMC Genomics *7*, 311. https://doi.org/10.1186/1471-2164-7-311.

85. Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T., and Ast, G. (2012). Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. Genome Res. *22*, 35–50. https://doi.org/10.1101/gr.119834.110.

86. De Conti, L., Baralle, M., and Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip. Rev. RNA *4*, 49–60. https://doi.org/10.1002/wrna.1140.

87. Schneider, M., Will, C.L., Anokhina, M., Tazi, J., Urlaub, H., and Lührmann, R. (2010). Exon definition complexes contain the tri-snRNP and can be directly converted into B-like precatalytic splicing complexes. Mol. Cell *38*, 223–235. https://doi.org/10.1016/j.molcel.2010.02.027.

88. Sharma, S., Wongpalee, S.P., Vashisht, A., Wohlschlegel, J.A., and Black, D.L. (2014). Stem-loop 4 of U1 snRNA is essential for splicing and interacts with the U2 snRNP-specific SF3A1 protein during spliceosome assembly. Genes Dev. *28*, 2518–2531. https://doi.org/10.1101/gad.248625.114.

89. Martelly, W., Fellows, B., Senior, K., Marlowe, T., and Sharma, S. (2019). Identification of a noncanonical RNA binding domain in the U2 snRNP protein SF3A1. RNA *25*, 1509–1521. https://doi.org/10.1261/rna.072256.119.

90. Plaschka, C., Lin, P.-C., Charenton, C., and Nagai, K. (2018). Prespliceosome structure provides insights into spliceosome assembly and regulation. Nature *559*, 419–422. https://doi.org/10.1038/s41586-018-0323-8.

91. Martelly, W., Fellows, B., Kang, P., Vashisht, A., Wohlschlegel, J.A., and Sharma, S. (2021). Synergistic roles for human U1 snRNA stem-loops in pre-mRNA splicing. RNA Biol. *18*, 2576–2593. https://doi.org/10.1080/15476286.2021.1932360.

92. Linares, A.J., Lin, C.-H., Damianov, A., Adams, K.L., Novitch, B.G., and Black, D.L. (2015). The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. ELife *4*, e09268. https://doi.org/10.7554/eLife.09268.

93. Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR *6*, 277–293. https://doi.org/10.1007/BF00197809.

94. Lee, W., Tonelli, M., and Markley, J.L. (2015). NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics *31*, 1325–1327. https://doi.org/10.1093/bioinformatics/btu830.

95. Güntert, P. (2009). Automated structure determination from NMR spectra. Eur. Biophys. J. *38*, 129–143. https://doi.org/10.1007/s00249-008-0367-z.

96. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009). TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J. Biomol. NMR *44*, 213–223. https://doi.org/10.1007/s10858-009-9333-z.

97. Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T.E., and Nilges, M. (2007). ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics *23*, 381–382. https://doi.org/10.1093/bioinformatics/btl589.

98. Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). Aqua and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR *8*, 477–486. https://doi.org/10.1007/BF00228148.

99. Bhattacharya, A., Tejero, R., and Montelione, G.T. (2007). Evaluating protein structures determined by structural genomics consortia. Proteins *66*, 778–795. https://doi.org/10.1002/prot.21165.

100. Koradi, R., Billeter, M., and Wüthrich, K. (1996). MOLMOL: A program for display and analysis of macromolecular structures. J. Mol. Graph. *14*, 51–55. https://doi.org/10.1016/0263-7855(96)00009-4.

101. Schrödinger, L., and DeLano, W. (2020). PyMOL. http://www.pymol.org/pymol.

102. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682. https://doi.org/10.1038/nmeth.2019.

103. Coleman, T., Branch, M.A., and Grace, A. (1999). Optimization Toolbox. For Use with MATLAB. User's guide. The MathWorks Inc, Ver. 2.

104. R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org/.

105. Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., González-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. ELife *5*, e11752. https://doi.org/10.7554/eLife.11752.

106. Dosch, J., Bergmann, H., Tran, V., and Ebersberger, I. (2023). FAS: assessing the similarity between proteins using multi-layered feature architectures. Bioinformatics *39*, btad226. https://doi.org/10.1093/bioinformatics/btad226.

107. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

108. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

109. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li,

H. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

110. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

111. Roehr, J.T., Dieterich, C., and Reinert, K. (2017). Flexbar 3.0 - SIMD and multicore parallelization. Bioinformatics *33*, 2941–2942. https://doi.org/10.1093/bioinformatics/btx330.

112. Krakau, S., Richard, H., and Marsico, A. (2017). PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. Genome Biol. *18*, 240. https://doi.org/10.1186/s13059-017-1364-2.

113. Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA package 2.0. Algorithms Mol. Biol. *6*, 26. https://doi.org/10.1186/1748-7188-6-26.

114. Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014). iCLIP: protein–RNA interactions at nucleotide resolution. Methods *65*, 274–287. https://doi.org/10.1016/j.ymeth.2013.10.011.

115. Spellman, R., Llorian, M., and Smith, C.W.J. (2007). Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. Mol. Cell *27*, 420–434. https://doi.org/10.1016/j.molcel.2007.06.016.

116. Coelho, M.B., Attig, J., Bellora, N., König, J., Hallegger, M., Kayikci, M., Eyras, E., Ule, J., and Smith, C.W.J. (2015). Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. EMBO J. *34*, 653–668. https://doi.org/10.15252/embj.201489852.

117. Grzesiek, S., and Bax, A. (1992). Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. J. Am. Chem. Soc. *114*, 6291–6293. https://doi.org/10.1021/ja00042a003.

118. Sattler, M., Schleucher, J., and Griesinger, C. (1999). Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog. Nucl. Magn. Reson. Spectrosc. *34*, 93–158. https://doi.org/10.1016/s0079-6565(98)00025-9.

119. Wishart, D.S., and Sykes, B.D. (1994). The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. J. Biomol. NMR *4*, 171–180. https://doi.org/10.1007/BF00175245.

120. Saitô, H. (1986). Conformation-dependent ^{13}C chemical shifts: a new means of conformational characterization as obtained by high-resolution solid-state ^{13}C NMR. Magn. Reson. Chem. *24*, 835–852. https://doi.org/10.1002/mrc.1260241002.

121. Kjaergaard, M., and Poulsen, F.M. (2011). Sequence correction of random coil chemical shifts: correlation between neighbor correction factors and changes in the Ramachandran distribution. J. Biomol. NMR *50*, 157–165. https://doi.org/10.1007/s10858-011-9508-2.

122. Farrow, N.A., Muhandiram, R., Singer, A.U., Pascal, S.M., Kay, C.M., Gish, G., Shoelson, S.E., Pawson, T., Forman-Kay, J.D., and Kay, L.E. (1994). Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. Biochemistry *33*, 5984–6003. https://doi.org/10.1021/bi00185a040.

123. Mulder, F.A., Schipper, D., Bott, R., and Boelens, R. (1999). Altered flexibility in the substrate-binding site of related native and engineered high-alkaline *Bacillus subtilisins*. J. Mol. Biol. *292*, 111–123. https://doi.org/10.1006/jmbi.1999.3034.

124. Williamson, M.P. (2013). Using chemical shift perturbation to characterise ligand binding. Prog. Nucl. Magn. Reson. Spectrosc. *73*, 1–16. https://doi.org/10.1016/j.pnmrs.2013.02.001.

125. Zwahlen, C., Gardner, K.H., Sarma, S.P., Horita, D.A., Byrd, R.A., and Kay, L.E. (1998). An NMR experiment for measuring methyl-methyl NOEs in ^{13}C-labeled proteins with high resolution. J. Am. Chem. Soc. *120*, 7617–7625. https://doi.org/10.1021/ja981205z.

126. Marsh, J.A., Singh, V.K., Jia, Z., and Forman-Kay, J.D. (2006). Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. Protein Sci. *15*, 2795–2804. https://doi.org/10.1110/ps.062465306.

127. Linge, J.P., Williams, M.A., Spronk, C.A.E.M., Bonvin, A.M.J.J., and Nilges, M. (2003). Refinement of protein structures in explicit solvent. Proteins *50*, 496–506. https://doi.org/10.1002/prot.10299.

128. Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921. https://doi.org/10.1107/s0907444998003254.

129. Messias, A.C., and Sattler, M. (2004). Structural basis of single-stranded RNA recognition. Acc. Chem. Res. *37*, 279–287. https://doi.org/10.1021/ar030034m.

130. Wiemann, S., Pennacchio, C., Hu, Y., Hunter, P., Harbers, M., Amiet, A., Bethel, G., Busse, M., Carninci, P., Dunham, I., et al. (2016). The ORFeome Collaboration: a genome-scale human ORF-clone resource. Nature Methods *13*, 191–192.

131. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*. D766–D773. https://doi.org/10.1093/nar/gky955.

132. Busch, A., Brüggemann, M., Ebersberger, S., and Zarnack, K. (2020). iCLIP data analysis: a complete pipeline from sequencing reads to RBP binding sites. Methods *178*, 49–62. https://doi.org/10.1016/j.ymeth.2019.11.008.

133. Paggi, J.M., and Bejerano, G. (2018). A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. RNA *24*, 1647–1658. https://doi.org/10.1261/rna.066290.118.

134. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC genome browser database: update 2006. Nucleic Acids Res. *34*. D590–D598. https://doi.org/10.1093/nar/gkj144.

135. Green, C.J., Gazzara, M.R., and Barash, Y. (2018). MAJIQ-SPEL: web-tool to interrogate classical and complex splicing variations from RNA-Seq data. Bioinformatics *34*, 300–302. https://doi.org/10.1093/bioinformatics/btx565.

136. Norton, S.S., Vaquero-Garcia, J., Lahens, N.F., Grant, G.R., and Barash, Y. (2018). Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. Bioinformatics *34*, 1488–1497. https://doi.org/10.1093/bioinformatics/btx790.

137. Zhang, J., Bajari, R., Andric, D., Gerthoffert, F., Lepsa, A., Nahal-Bose, H., Stein, L.D., and Ferretti, V. (2019). The International Cancer Genome Consortium data portal. Nat. Biotechnol. *37*, 367–369. https://doi.org/10.1038/s41587-019-0055-9.

138. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095.

139. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1. https://doi.org/10.1126/scisignal.2004088.

140. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P.,

et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

141. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*. D941–D947. https://doi.org/10.1093/nar/gky1015.

142. Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M. (2016). Toward a shared vision for cancer genomic data. N. Engl. J. Med. *375*, 1109–1112. https://doi.org/10.1056/NEJMp1607591.

143. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*. D1062–D1067. https://doi.org/10.1093/nar/gkx1153.

144. Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394. https://doi.org/10.1089/1066527041410418.

145. Birikmen, M., Bohnsack, K.E., Tran, V., Somayaji, S., Bohnsack, M.T., and Ebersberger, I. (2021). Tracing eukaryotic ribosome biogenesis factors into the archaeal domain sheds light on the evolution of functional complexity. Front. Microbiol. *12*, 739000. https://doi.org/10.3389/fmicb.2021.739000.

146. Smith, T., Heger, A., and Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. *27*, 491–499. https://doi.org/10.1101/gr.209601.116.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit anti-FUBP1	GeneTex	Cat# GTX104579; RRID: AB_11165485
Mouse anti-U2AF2	Sigma-Aldrich	Cat# U4758; RRID: AB_262122
Mouse anti-SF3B1	MBL	Cat# D221-3; RRID: AB_592712
Mouse anti-SF1	Abnova	Cat# H00007536-M01A; RRID: AB_10774630
rabbit anti-PTBP1	Christopher Smith	Linares et al.[92]
Mouse anti-vinculin	Sigma-Aldrich	Cat# V9264; RRID: AB_10603627
Goat anti-rabbit IgG, HRP-linked	Cell Signaling	Cat# 7074S; RRID: AB_2099233
Horse anti-mouse IgG, HRP-linked	Cell Signaling	Cat# 7076S; RRID: AB_330924
Bacterial and virus strains		
DH5alpha	Invitrogen	Cat# 18265017
MACH1	Invitrogen	Cat# C862003
E. coli BL21-CodonPlus (DE3)-RIL	Agilent	Cat# 230245
E. coli BL21 (DE3)	Sigma-Aldrich	Cat# CMC0014
Chemicals, peptides, and recombinant proteins		
FUGENE HD reagent	Promega	Cat# E2311
Lipofectamine CRISPRMAX reagent	Thermo Fisher	Cat# CMAX00001
Lipofectamine RNAimax	Thermo Fisher	Cat# 13778150
Lipofectamine 2000	Invitrogen	Cat# 11668019
cOmplete Protease-Inhibitor Mix	Sigma-Aldrich	Cat# 4693159001
TURBO DNase	Thermo Fisher	Cat# AM2238
SuperSignal West PICO Chemiluminescent Substrate	Thermo Fisher	Cat# 15626144
4-thiouridine	Sigma-Aldrich	Cat# T4509-25MG
T4 RNA ligase	New England Biolabs	Cat# M0202S
T4 RNA ligase 1	New England Biolabs	Cat# M0437M
pCp-Cy5	Jena Bioscience	Cat# NU-1706-CY5
T7 RNA polymerase	Geerlof A., Protein Expression and Purification Facility, HMGU Munich	N/A
Pfu DNA Polymerase	Promega	Cat# M7741
OneTaq DNA Polymerase	New England Biolabs	Cat# M0480S
Phusion High-Fidelity DNA Polymerase	New England Biolabs	Cat# M0530S
Critical commercial assays		
TranscriptAid Enzyme Mix	Thermo Fisher	Cat# K0441
GeneArt Genomic Cleavage Detection Assay	Thermo Fisher	Cat# A24372
Zero Blunt TOPO PCR Cloning Kit	Thermo Fisher	Cat# 451245
RNeasy PLUS Mini Kit	Qiagen	Cat# 74034
TruSeq library preparation Kit "Ribo-Zero Gold"	Illumina	Cat# 20040526
RevertAid First Strand cDNA Synthesis Kit	Thermo Fisher	Cat# 10161310
Q5 Site-Directed Mutagenesis Kit	New England Biolabs	Cat# E0552S
High Sensitivity D1000 ScreenTape	Agilent	Cat# 5067-5584
High Sensitivity RNA ScreenTape	Agilent	Cat# 5067-5579
NuPAGE 1 mm, 4-12% Bis-Tris Mini Protein Gel	Thermo FIsher	Cat# 12090156
HiScribe T7 High Yield RNA Synthesis Kit	New England Biolabs	Cat# E2040S

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ProNex Dual Size-Selective Purification System	Promega	Cat# NG2002
BP clonase II mix kit	Invitrogen	Cat# 10348582
LR clonase technology	Invitrogen	Cat# 11791020
Deposited data		
in vitro and *in vivo* iCLIP and RNA-Seq data	This study	GEO: GSE220186
Kinetic modeling of cassette exon splicing	This study	https://doi.org/10.5281/zenodo.8076768
Protein structure data	This study	PDB: 8P25
NMR data	This study	BMRB: 34816
Original Western blot, gel images and capillary electrophoresis images	This study, Mendeley Data	https://doi.org/10.17632/nj8ybm8vb2.1
RNA-Seq data: control and shRNA knockdown for FUBP1 in K562 cells	Luo et al.[63], ENCODE Project Consortium[62]	ENCODE: ENCSR260BQC (control) and ENCSR608IXR (FUBP1 KD)
Differentially spliced junctions in splicing factor mutations	Seiler et al.[1]	Table S3 in Seiler et al.
Experimental models: Cell lines		
human: HeLa	ATCC	Cat# CCL-2, RRID:CVCL_0030
human: RPE1 FUBP1 WT: hTERT-RPE1 NatNeo Cas9 Mono Puro sens	Manuel Kaulich	N/A
human: RPE1 FUBP1 KO: hTERT-RPE1 NatNeo Cas9 Mono Puro sens FUBP1 -/-	This study	N/A
human: RPE1 FUBP1 Nbox-mut: hTERT-RPE1 NatNeo Cas9 Mono Puro sens FUBP1 indel 31-40	This study	N/A
human: HEK293	DSMZ	ACC305
Oligonucleotides		
See Table S5	(too many oligos to list here)	N/A
Recombinant DNA		
See Table S6	(too many plasmids to list here)	N/A
Software and algorithms		
Topspin 3.5	Bruker	https://www.bruker.com/en/products-and-solutions/mr/nmr-software/topspin.html
NMRpipe	Delaglio et al.[93]	https://www.ibbr.umd.edu/nmrpipe/index.html
NMRFAM-Sparky	Lee et al.[94]	https://nmrfam.wisc.edu/nmrfam-sparky-distribution/
CYANA 3.98.13	Güntert[95]	https://cyana.org/wiki/Main_Page
TALOS+	Shen et al.[96]	https://spin.niddk.nih.gov/bax/software/TALOS/
ARIA2.3	Rieping et al.[97]	http://aria.pasteur.fr/
ProcheckNMR	Laskowski et al.[98]	https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/
PSVS	Bhattacharya et al.[99]	https://montelionelab.chem.rpi.edu/PSVS/PSVS/
MolMol	Koradi et al.[100]	https://sourceforge.net/p/molmol/wiki/Home/
PYMOL	Schrödinger and DeLano[101]	https://pymol.org/2/
ImageJ 2.1.0	Schindelin et al.[102]	https://imagej.net/
MicroCalPEAQ ITC Analysis software	Malvern Panalytical	https://www.malvernpanalytical.com/
Agilent TapeStation Software 5.1	Agilent	https://www.agilent.com
Image Lab 6.0.1 build 34	bio-rad	https://www.bio-rad.com/
MATLAB	Coleman et al.[103]	https://www.mathworks.com/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R 4.1.1.	Core Team[104]	https://www.r-project.org/
MAJIQ v2.3	Vaquero-Garcia et al.[105]	https://majiq.biociphers.org/
FAS	Dosch et al.[106]	https://github.com/BIONF/FAS
fDOG	N/A	https://github.com/BIONF/fDOG
STAR	Dobin et al.[107]	https://github.com/alexdobin/STAR
Cutadapt 2.4	Martin[108]	https://cutadapt.readthedocs.io/en/stable/
Samtools v1.9	Danecek et al.[109]	http://www.htslib.org/
Subread tool suite v1.6.2	Liao et al.[110]	https://subread.sourceforge.net/
FastQC v0.11.8	N/A	https://www.bioinformatics.babraham.ac.uk/projects/fastqc
FASTX-Toolkit v0.0.14	N/A	http://hannonlab.cshl.edu/fastx_toolkit/
seqtk v1.3	N/A	https://github.com/lh3/seqtk/
Flexbar v3.4.0	Roehr et al.[111]	https://github.com/seqan/flexbar
PureCLIP v1.3.1	Krakau et al.[112]	https://github.com/skrakau/PureCLIP
ViennaRNA Package 2.4.17	Lorenz et al.[113]	https://www.tbi.univie.ac.at/RNA/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Julian König (j.koenig@imb-mainz.de).

Materials availability

All unique/stable reagents generated in this study are available from the lead contact.

Data and code availability

- RNA-seq, *in vivo* and *in vitro* iCLIP data have been deposited at GEO and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. Protein structures have been deposited to the Protein Data Bank and are available under the accession number 8P25. NMR data used for structure calculation are deposited in the BMRB under the accession code 34816. Original Western blot, gel images and capillary electrophoresis images have been deposited at Mendeley Data and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- This paper analyses existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at GitHub and is publicly available as of the date of publication at https://doi.org/10.5281/zenodo.8076768.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

RPE1 cell lines and culture conditions

The hTERT-RPE1 NatNeo Cas9 Mono Puro sens cell line was a generous gift of the Kaulich lab at the Frankfurt CRISPR/Cas Screening Center (FCSC) and are modified from original hTERT RPE1 cells (ATCC, CRL-4000). Cells were grown and maintained in Dulbecco's modified Eagle's medium (DMEM): Nutrient Mixture F-12 (DMEM/F-12; Thermo Fisher 11530566), supplemented with 10% fetal bovine serum (PAN-Biotech), 2 mM glutamine (Thermo Fisher), 1% penicillin–streptomycin (Thermo Fisher), and 20 μg/ml hygromycin B (Thermo Fisher). Cells were incubated at 37°C with 5% CO_2. Subcultivation was performed with 3 ml of 0.1% trypsin every 2–3 days for 20 passages. After that, new cells were thawed from stocks containing 1×10^6 cells in 1 ml of growth medium, supplemented with 10% DMSO and 50% fetal bovine serum (FBS). For semi-quantitative RT-PCR, 1×10^5 RPE1 cells were seeded into one well of a six-well plate (Falcon), one day prior to transfection. DNA (2 μg) was diluted in 100 μl of OptiMEM and transfected with 6.4 μl of Fugene HD reagent (Promega). Cells were incubated at 37°C with 5% CO_2 for 24 h before harvesting. For RNA-seq, 1.5×10^6 cells were seeded in a 10-cm cell culture dish (Corning) 48 h prior to isolation.

HeLa cell line and culture conditions

HeLa cells (ATCC CCL-2) were grown and maintained in DMEM (Thermo Fisher), supplemented with 10% FBS, 2 mM glutamine (Thermo Fisher) and 1% penicillin–streptomycin (Thermo Fisher). Cells were incubated at 37°C with 5% CO_2. Subcultivation was performed with 3 ml of 0.1% trypsin every 2–3 days for 20 passages. After that, new cells were thawed from stocks containing 1×10^6 cells in 1 ml of growth medium, supplemented with 10% DMSO (Sigma) and 50% FBS.

HEK cell line and culture conditions

HEK293 cells (DSMZ) were grown and maintained in DMEM (Thermo Fisher), supplemented with 10% fetal bovine serum (PAN-Biotech), 2 mM glutamine (Thermo Fisher) and 1% penicillin–streptomycin (Thermo Fisher). Cells were incubated at 37°C with 5% CO_2. Subcultivation was performed with 1 ml of 0.05% trypsin every 2–3 days for up to 15 passages. Then, new cells were thawed from stocks containing 2×10^6 cells in 1 ml of growth medium, supplemented with 10% DMSO (Sigma) and 90% FBS.

Recombinant protein expression

Proteins were expressed in *E. coli* BL21 (DE3) cells grown in LB medium or M9 minimal medium supplemented with 1 g/l $^{15}NH_4Cl$ and 2 g/l ^{13}C-glucose (uniformly labeled) at 37°C. Protein expression was induced with 1.0 mM isopropyl β-D-1-thiogalactopyranoside (IPTG).

METHOD DETAILS

Establishing *FUBP1* KO/*Nbox*^{mut} cell lines

FUBP1 was mutated and knocked out using the CRISPR/Cas9 system in hTERT-RPE1 NatNeo mono puro sens cells. This cell line is puromycin sensitive and expresses *Streptococcus pyogenes* Cas9 under neomycin resistance. For the creation of the *FUBP1* KO and *FUBP1-Nbox*^{mut} RPE1 cell lines, cells were cultured as described above with the addition of neomycin (G418, InvivoGen) to preserve Cas9 expression. Guide RNA (gRNA) was amplified from oligos #54 and #55 (Table S5) with Phusion Polymerase (New England Biolabs) and *in vitro* transcribed with TranscriptAid Enzyme Mix (Thermo Fisher) according to the manufacturer's protocol. Cells were then transfected with the resulting gRNA using Lipofectamine CRISPRMAX (Thermo Fisher) according to the manufacturer's protocol and incubated for 48 h. To assess the general editing efficiency, a GeneArt Genomic Cleavage Detection Assay (Thermo Fisher) was performed. Edited cells were then sorted by fluorescence-activated cell sorting (FACS), and each cell was cultured in a separate well of a 96-well plate (Corning). From each clonal cell line, genomic DNA (gDNA) was isolated and amplified by PCR. The successful disruption of the targeted site was validated by enzyme restriction and Sanger sequencing (StarSEQ GmbH, Mainz, Germany) of the colonies. To obtain the novel sequence of the targeted site on both alleles, gDNA was also cloned into TOPO vectors using the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher), and the obtained plasmids were Sanger-sequenced. All Sanger sequencings were performed with oligo #56 (Table S5). The edited sequences led to mutated protein products, as shown in Figure S5G.

Immunoblotting

For each hTERT RPE1-derived cell line, 1×10^6 cells were seeded on a 10-cm cell culture dish (Corning) and harvested after incubation for 48 h at 37°C, 5% CO_2. Cells were lysed in modified RIPA buffer containing 50 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, 1% NP-40 (Sigma), 0.1% sodium deoxycholate (Sigma) and supplemented with cOmplete Protease Inhibitor Mix (Sigma), and TURBO DNase (Thermo Fisher) for 15 min on ice. Cell debris was precipitated by centrifugation at 16,000 ×*g* for 15 min at 4°C. The cleared protein lysate was transferred into a new reaction tube (Eppendorf) and the concentration was measured with a BCA Protein Assay Kit (Thermo Fisher). 20 μg of protein lysate was mixed with 4× NuPAGE LDS Sample Buffer and heated to 70°C for 10 min. Samples were loaded onto a NuPAGE 1 mm, 4–12% Bis-Tris Mini Protein Gel (Thermo Fisher) and electrophoresis was performed at 180 V, 400 mA for 50 min on a NuPAGE Novex Gel System (Invitrogen). Protein transfer to a nitrocellulose membrane (VWR International) was performed at 30 V, 400 mA over 60 min using the same gel system. The membrane was blocked in 5% milk diluted in PBS-T. The primary antibody (key resources table) was incubated overnight at 4°C, and the secondary antibody was incubated for 60 min at room temperature. All antibodies were diluted in 5% milk–PBS-T. Between blocking and primary and secondary antibody steps, the membrane was washed three times with PBS-T. Detection was performed with SuperSignal West PICO Chemiluminescent Substrate (Thermo Fisher) and BioRad GelDoc (BioRad).

RPE1 RNA-seq

For RPE1 RNA sequencing (ID: imb_koenig_2020_12) and semi-quantitative RT-PCR analysis, RPE1 cells were grown as described above. Cells were washed once with DPBS and harvested with a cell scraper in l ml of DPBS. Suspensions were centrifuged at 1,000 ×*g* for 1 min at 4°C. RNA was isolated from cell pellets using an RNeasy PLUS Mini Kit (Qiagen) according to the manufacturer's protocol. For sequencing, RNA concentration was measured by Qubit RNA BR Assay and integrity of the RNA was confirmed by Bioanalyzer RNA Nano Assay (Agilent). Ribosomal RNA was removed and the remaining RNA was reverse transcribed into cDNA using the TruSeq library preparation kit with Ribo-Zero Gold (Illumina). The libraries were sequenced on an Illumina NextSeq 500 sequencer as 159-nt single-end reads.

HeLa RNA-seq

200,000 cells were seeded per well in a six-well dish 24 h prior to siRNA treatment. RNA-seq to assess intron splicing in HeLa cells (ID: imb_koenig_2018_18) was performed in four replicates. HeLa cells underwent a control knockdown (KD) with no-target siRNA. Oligos #40–#43 (Table S5) were delivered into cells using 3 μl of Lipofectamine RNAimax (Thermo Fisher) in 100 μl of OptiMEM to achieve a final siRNA concentration of 20 nM. Cells were harvested after incubation for 48 h. RNA was isolated from cell pellets using RNeasy PLUS Mini Kit (Qiagen) according to the manufacturer's protocol. RNA concentration was measured by Qubit RNA BR Assay, and integrity of the RNA was confirmed by Bioanalyzer RNA Nano Assay (Agilent). Ribosomal RNA was removed and the remaining RNA was reverse transcribed into cDNA using the TruSeq library preparation kit with Ribo-Zero Gold (Illumina). RNA-seq samples were sequenced on an Illumina NextSeq 500 sequencer as 84-nt single-end reads.

Semi-quantitative RT-PCR

The *MPDZ* minigene was created from HeLa gDNA extracts by amplification of chr9:13,183,353-13,189,041 with Phusion HighFidelity Polymerase (New England Biolabs). The PCR fragment was cloned into a pCR2.1 vector by Gibson assembly (IMB Protein Production Core Facility). *MPDZ* introns were shortened using a Q5 Site-Directed Mutagenesis Kit (New England Biolabs), resulting in $MPDZ^{\Delta intron}$, which lacks chr9:13,186,637-13,188,633 and chr9:13,183,736-13,186,120, $MPDZ^{\Delta BS}$, lacking chr9:13,186,494-13,186,618 and chr9:13,183,632-13,186,718, and $MPDZ^{\Delta intron+\Delta BS}$, lacking chr9:13,186,494-13,188,633 and chr9:13,183,632-13,186,120 (Figure S6B). The open reading frames for GFP and the FUBP1 variants ($FUBP1^{FL}$, $FUBP1^{\Delta N}$, $FUBP1^{A38D}$, $FUBP1^{\Delta C}$, $FUBP1^{W586,615R}$) used in the complementation assay were integrated in a pcDNA5 vector containing a CMV promoter and an N-terminal GFP tag, which was then used to transform DH5alpha cells (Invitrogen). All expression vectors and minigenes are described in Table S6. Plasmid purification was performed with the Qiaprep Spin Miniprep Kit (Qiagen) or the Qiaprep Plasmid Plus Midi Kit (Qiagen). Sequences were verified by Sanger sequencing. All hTERT RPE1 cell lines were seeded, transfected, and harvested as described in the section "RPE1 cell culture". For complementation, an equimolar amount of expression vector and minigene was used. RNA was isolated with the RNeasy Plus Mini Kit (Qiagen) and reverse transcribed using the RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher). The minigene cDNA was then amplified using OneTaq DNA Polymerase according to the manufacturer's protocol and oligos #57 and #58 as primers (Table S5). Splicing products were assessed on a High Sensitivity D1000 ScreenTape (Agilent) (Figure S6D). The percent spliced-in (PSI) value for the alternative exon was determined using the following formula: *Inclusion / (Inclusion + Skipping)*. PSI values in the complementation experiment were normalized to the mean of the wild-type (WT) within each condition. Statistical significance was assessed by Student's t-test and multiple testing correction was performed using the false discovery rate (FDR).

In vivo iCLIP

In vivo iCLIP was used to study protein–RNA interactions with individual nucleotide resolution.[43] For the U2AF2 *in vivo* iCLIP study, data from two iCLIP experiments were combined. The first U2AF2 and PTBP1 *in vivo* iCLIP experiments were performed as previously described.[114] The second U2AF2 *in vivo* iCLIP experiment as well as *in vivo* iCLIP experiments on FUBP1, SF1, and SF3B1 were performed using the iCLIP2 protocol as previously described.[44] In brief, HeLa cells were irradiated (150 mJ/cm^2) in a CL1000 UV crosslinker (UPV) to covalently bond the RNA-binding proteins to the bound nucleic acids. For *in vivo* iCLIP of FUBP1, crosslinking was achieved by 4-thiouridine (4sU)-mediated crosslinking (see section below). During subsequent cell lysis, the lysate was DNase-treated with TURBO DNase (Thermo Fisher) and RNA was partially digested to create 50–200-nt fragments. Immunoprecipitation of the investigated proteins was performed with antibodies listed in the key resources table. The anti-PTBP1 antibody was a kind gift from Christopher Smith.[115] Radioactive labeling at the 3′ end of the precipitated RNA enables visualization of the RNP complex by SDS-PAGE and transfer to a nitrocellulose membrane. After recovery of protein–RNA complexes from the membrane, proteinase K digestion resulted in protein-free RNA. cDNA was synthesized by reverse transcription, which stops at the crosslinked site, leading to truncated reads in the sequencing. The cDNA was cleaned twice using MyONE Silane beads (Thermo Fisher). PCR amplification and ProNex size selection were performed to amplify and purify the library, respectively. *In vivo* iCLIP libraries (except PTBP1 libraries) were sequenced on an Illumina NextSeq 500 sequencer as 92-nt single-end reads including a 6-nt (or 4-nt in the case of the first U2AF2 iCLIP) sample barcode as well as 5+4-nt (or 3+2-nt) unique molecular identifiers (UMIs). PTBP1 iCLIP libraries were sequenced on an Illumina GA-II machine[116] and then re-sequenced on an Illumina HiSeq 2000 machine as 50-nt single-end reads including a 4-nt sample barcode and 3+2-nt UMIs.

4-thiouridine crosslinking of FUBP1 in vivo *iCLIP*

For the FUBP1 *in vivo* iCLIP, HeLa cells were 4sU-labeled by adding 0.1 M 4sU in DMSO to a final concentration of 100 μM in a 10-cm cell culture dish. Cells were incubated for 16 h at 37°C, 5% CO_2, with the exclusion of light. After incubation, the cells were moved onto ice, shielded from light and irradiated at 365 nm, 800 mJ. Then, iCLIP was performed as described above.

In vitro iCLIP

In vitro iCLIP measures the intrinsic RNA-binding affinity of an RNA-binding protein (RBP).[28] To that end, recombinant proteins and *in vitro* transcripts resembling long natural transcripts[28] or a large-scale RNA pool transcribed from an oligonucleotide library were mixed and subjected to UV crosslinking and immunoprecipitation of the RBP of interest.

Production of recombinant proteins

N-terminally 6xHis-tagged U2AF2^{RRM12} was purified as previously described.[28] In brief, a recombinant construct (Table S6) was expressed in *E. coli* BL21-CodonPlus (DE3)-RIL cells (Agilent) for 3–4 h at 37°C using LB-Media and 1 mM IPTG. U2AF2^{RRM12} was purified using Ni Sepharose 6 Fast Flow beads (GE Healthcare) according to the manufacturer's protocol, and concentrated with Spin-X UF 500 5K MWCO columns (Corning) to a concentration of 1.156 mg/ml before being flash-frozen in liquid nitrogen and stored at −80°C. All three N-terminally 6xHis-tagged FUBP1 protein variants (FUBP1^{FL}, FUBP1$^{\Delta N}$, FUBP1^{N74}; Table S6) were expressed overnight at 16°C using LB media and 1 mM IPTG. Cells were lysed in lysis buffer (50 mM Tris-Cl, pH 8.0, 500 mM NaCl, 1 mM DTT, 5% glycerol, EDTA-free cOmplete protease inhibitor cocktail), using a CF1 Cell Disrupter (Constant Systems). Lysates were cleared by centrifugation (40,000 ×*g*, 30 min, 4°C). Recombinant proteins were affinity-purified from cleared lysates using an NGC Quest Plus FPLC system (Biorad) and a HisTrap FF 5 ml column (Cytiva) according to the manufacturers' protocols. Full-length FUBP1^{FL} and FUBP1$^{\Delta N}$ proteins were diluted 1:10 in heparin binding buffer (30 mM Na-HEPES, 20 mM NaCl, 5% glycerol, 1 mM DTT, pH 7.4), loaded onto a Heparin HP 5 ml column (Cytiva) and eluted over 15 column volumes using a linear gradient of 20–1000 mM NaCl in the heparin binding buffer. All FUBP1 variants were concentrated using Amicon 15 ml spin concentrators (Merck Millipore) and subjected to gel filtration (Superdex 200 16/60 pg in 30 mM Na-HEPES, 100 mM NaCl, 1 mM DTT, 5% glycerol, pH 7.4). Peak fractions containing the recombinant proteins after gel filtration were pooled, and protein concentration was determined by using absorbance spectroscopy and the respective extinction coefficient at 280 nm, before aliquots were flash-frozen in liquid nitrogen and stored at −80°C. For the detailed workflow, log files can be requested from Dr. Julian König.

Preparation of *long* in vitro *transcripts*

Long *in vitro* transcripts were prepared as described in Sutandy et al.[28] Minigene and spike-in RNAs were created by PCR amplification of DNA templates using Phusion High-Fidelity DNA Polymerase (New England Biolabs) according to the manufacturer's protocol. *In vitro* transcription of gel-purified PCR products was performed using HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs) according to the manufacturer's instructions. RNA was isolated with the RNeasy Plus Mini Kit (Qiagen), followed by DNA digestion with TURBO DNase and another RNA extraction. RNA quality was verified by capillary electrophoresis using High Sensitivity RNA ScreenTape (Agilent). RNA concentration was measured with a Qubit RNA HS Assay Kit (Thermo Fisher). Aliquots of equimolar mixes of all minigenes as well as spike-in aliquots were stored at −80°C.

In vitro *iCLIP with long* in vitro *transcripts*

In vitro iCLIP with long *in vitro* transcripts (ID: imb_koenig_2018_01_sub16) was performed for U2AF2^{RRM12} alone or supplemented with different FUBP1 variants. The experiment was performed with a pool of eight *in vitro* transcripts (*C4BPB*, *MPDZ*, *MYC*, *MYL6*, *NF1*, *TENT2*, *PCBP2*, and *PTBP2*, see GEO record GSE220183) as previously described.[28] The *in vitro* transcripts were preheated for 5 min at 70°C to minimize RNA secondary structure. Then, *in vitro* transcripts at a final concentration of 2 nM were added to 50 nM U2AF2^{RRM12} either alone (three replicates) or supplemented with either 50 nM FUBP1^{FL} (two replicates), 50 nM FUBP1$^{\Delta N}$ (two replicates), or 50 nM FUBP1^{N74} (two replicates). The mixtures were incubated at 37°C for 5 min before UV irradiation at 50 mJ/cm^2. The *in vitro* iCLIP reaction was spiked with 10 μl of crosslinked mixture containing 250 nM U2AF2^{RRM12} and 6 nM *NUP133 in vitro* transcript for normalization.[28] Partial RNase digestion and DNase treatment, followed by the standard iCLIP protocol, were performed as described in the section "*In vivo iCLIP*". After reverse transcription, the cDNA was purified and libraries were generated according to the iCLIP2 protocol.[44]

Preparation of oligo-derived transcripts

A total of 1,998 DNA oligonucleotides were chosen to represent 182-nt regions around 3′ splice sites, including the last 132 nt upstream of a 3′ splice site and the first 50 nt of the downstream exon, preceded by 18 nt of T7 promoter sequence for the reverse transcription. The genomic coordinates of all regions represented in the oligonucleotide library are listed in GEO record GSE220183. The DNA oligonucleotides were purchased from TWIST Bioscience (South San Francisco, CA). Before *in vitro* transcription, L3 adapter ligation was performed. This was achieved by resuspending the DNA pellet in T4 RNA ligase (New England Biolabs) mix containing a 1:10 oligo/adapter ratio for high ligation efficiency. This mixture was reacted overnight at 16°C at 1300 rpm and then inactivated at 98°C for 5 min. L3-ligated DNA oligonucleotide (2.6 ng) was amplified using the Phusion High-Fidelity DNA Polymerase (New England Biolabs) according to the manufacturer's protocol. Amplicons were purified twice using the ProNex Dual Size-Selective Purification System (Promega) with an optimized bead/library ratio of first 1.13 and then 0.5. Capillary electrophoresis with a High Sensitivity D1000 ScreenTape (Agilent) was used for quality control. Then, *in vitro* transcription was performed for 4 h at 37°C by following the HiScribe T7 (New England Biolabs) protocol for short transcripts. Subsequently, RNA was treated with TURBO DNase I and isolated using Qiagen's protocol for "Total RNA containing small RNA from cells" (RNeasy Plus Mini Handbook, Appendix E) with the reagents mentioned above.

in vitro *iCLIP on oligo-derived transcripts*

For *in vitro* iCLIP with an oligonucleotide-derived RNA pool (ID: imb_koenig_2018_01_sub12), the oligonucleotide-derived transcript pool at a concentration of 50 nM was preheated for 5 min at 70°C and incubated with 50 nM U2AF2^{RRM12} alone or with either 50 or 300 nM FUBP1^{FL} (three replicates each) for 10 min before UV irradiation at 50 mJ/cm^2. iCLIP was performed as described in the section "*In vivo iCLIP*", omitting the partial RNase digestion and L3 linker ligation steps as they do not apply here. The reaction was spiked with a mix of 10 150-nt long spike-in oligonucleotides for normalization (oligos #44–#53; Table S5).

Sequencing and data preprocessing

In vitro iCLIP libraries were sequenced on an Illumina NextSeq 500 sequencer as 150-nt single-end reads including a 6-nt sample barcode as well as 5+4-nt UMIs. The reads were bioinformatically preprocessed as described for *in vivo* iCLIP samples. The number of uniquely mapped reads for all *in vitro* iCLIP samples are given in Table S1.

Protein expression and purification

All plasmids encoding sequences of FUBP1, U2AF2, chimeric U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-}box}$ (linked by a 14 GS linker), SF1, SNRPA, SNRPB, and PRPF40B were cloned into the pETM11 vector or pET24 vector with a His tag, His-GB1 tag, or His-protein A tag, followed by a TEV cleavage site. The point mutants of FUBP1 were generated by site-directed mutagenesis. All constructs are listed in Table S6.

Recombinant proteins were expressed in *E. coli* BL21 (DE3) cells in LB medium or M9 minimal medium supplemented with 1 g/l $^{15}NH_4Cl$ and 2 g/l ^{13}C-glucose (uniformly labeled). After growth of the bacterial cells to an OD_{600} value of 0.8, protein expression was induced with 1.0 mM IPTG followed by overnight expression at 18°C. After resuspension in 50 mM Tris, pH 8.0, 500 mM NaCl, 10 mM imidazole (supplemented with lysozyme, 1 mg/ml DNase, 2 mM $MgSO_4$, and protease inhibitor), the cells were lysed using a French press. Cleared lysates were added to Ni–NTA resin, washed with 2 M NaCl and eluted with 500 mM imidazole. The His tag was cleaved with His-tagged TEV protease at 4°C overnight. The protein was further purified by removing the cleaved His tag, uncleaved protein and TEV protease from the desired protein on a second Ni–NTA column. All proteins were further purified by ion-exchange chromatography on RESOURCE S or RESOURCE Q columns (Cytiva) (20 mM Tris, pH 8.0 or 20 mM sodium phosphate, pH 6.5, gradient from 0 to 1 M NaCl in 10 column volumes) followed by size-exclusion chromatography on a HiLoad 16/600 Superdex 75 column (GE Healthcare) (20 mM sodium phosphate, pH 6.5, 150 mM NaCl).

NMR spectroscopy

All NMR samples ($^{13}C^{15}N$- or ^{15}N-labeled, as appropriate) were measured at concentrations of 0.1–1 mM in NMR buffer (20 mM sodium phosphate, pH 6.5, 50 mM NaCl, 2 mM DTT) containing 10% (v/v) D_2O at 25°C on 900-, 800-, 600-, or 500-MHz Bruker Avance NMR spectrometers (cryogenic triple-resonance gradient probes). The NMR spectra were processed with TOPSPIN3.5 (Bruker) or NMRPipe[93] and analyzed using NMRFAM-Sparky.[94]

Chemical shift assignment

Protein backbone assignments were obtained from standard HNCA, HNCACB, CBCA(CO)NH, HNHA backbone experiments. Specifically, for KH domains, the ^1H–^{15}N HSQC spectrum of KH1–4 was first assigned, then corresponding assignments were transferred to the spectra of the individual and tandem KH domains. Further side-chain resonances were assigned using CC(CO)NH, HCC(CO)NH, hCCH-TOCSY and HcCH-TOCSY experiments. The distance restraints for structure calculations were obtained from 3D ^{15}N- and ^{13}C-edited NOESY–HSQC experiments.[117,118] Secondary structure propensities were derived from the difference of C_a and C_b chemical shifts to the random coil shifts.[119–121]

Relaxation experiments

^{15}N-relaxation experiments were recorded on an 800 MHz Bruker Avance NMR spectrometer at 25°C and ^{15}N T_1 and T_2 relaxation times were acquired from pseudo-3D HSQC experiments in an interleaved manner with eight relaxation delays for T_1 (20, 60, 100, 200, 400, 600, 800, 1200 ms) and nine relaxation delays for T_2 (16.96, 33.92, 67.84, 101.76, 135.68, 169.6, 254.4, 305.28, 339.2 ms).[122] Residual relaxation rates were obtained by fitting the data to an exponential function using NMRFAM-Sparky.[94]

Titrations

For NMR titrations, ^1H–^{15}N HSQC spectra were measured after each addition of titrant and the changes were visualized by calculating the CSP.[123] The K_D values were calculated from NMR titrations by plotting the CSP of selected peaks (8) against the ligand concentration and fitting the data as previously described. Standard deviations of the mean were calculated from K_D values of the 8 selected peaks.[124]

Structure calculation

To stabilize the U2AF2 and FUBP1 interaction, a chimeric construct of U2AF2^{RRM2} and FUBP1$^{N\text{-}box}$ was introduced for the subsequent structure determination (Table S6). Overall structural integrity of the chimeric construct and recapitulation of the interaction was confirmed by comparing ^1H–^{15}N HSQC spectra of the chimeric construct to that of the intermolecular complex U2AF2-RRM2–FUBP1$^{N\text{-}box}$ (Figures S3I and S3J). CYANA3 (3.98.15) was used for automated NOE assignments and initial structure calculations.[95] To overcome partial signal broadenings for the resonances at the interface of the two domains, possibly due to the weaker affinity, additional unambiguous intramolecular distance restraints from ^{13}C-NOESY–HMQC and methyl-NOESY spectra were manually assigned and included in the structure calculation.[125] A minimal number of typical hydrogen bonds, which were confirmed by ^{15}N-edited NOESY and secondary structure propensity, was implemented to assist the initial folding during the structure calculation. Dihedral angle restraints were derived from SSP and ^{13}C secondary chemical shifts using TALOS+, including resonances of Ca, Cb, C, H, and N.[96,126] For water refinement, distance restraints from CYANA3 considering an error of ± 0.5 Å are used. Water refinement[127] of the 20 lowest-energy structures (500 initial structures) was performed with ARIA2.3[97] and CNS.[128] The quality of the 10 final structures was evaluated by ProcheckNMR[98] and PSVS.[99] Ensemble structure root mean square (r.m.s.) deviations were calculated using MolMol[100] and the ribbon representations were prepared in PyMOL (The PyMOL Molecular Graphics System, version 1.8.6.0, Schrödinger, LLC). Structural statistics are shown in Table 1.

Scaffold-independent analysis

For the initial screening, the 16 DNA pools of 5-mer DNA (Table S5, #63, IDT), instead of RNA due to their similarity in binding, were generated by introducing a specific nucleotide at a designated position while randomizing the other four positions. Titrations of 100 μM FUBP1 KH domain samples with the different DNA pools (0.5, 1.0, 2.0, and 4.0 molar equivalents of titrant to analyte) were performed at 25°C in NMR buffer (20 mM sodium phosphate, pH 6.5, 50 mM NaCl, 2 mM DTT) containing 10% (v/v) D_2O by recording SOFAST HMQC spectra on a 600 MHz Bruker Avance NMR spectrometer (cryogenic triple-resonance gradient probe). For the comparison and identification of position-specific nucleotide preference, we focused on a subset of 12 representative peaks, which show visibly clear changes in chemical shift (fast-exchange regime) and are therefore involved in binding, for further analysis. CSPs of these peaks were calculated (see above) and the average CSPs of all peaks for each pool were normalized against the largest CSP calculated in the four pools to obtain a score for nucleotide preference at a specific position. The final optimized motifs were verified by comparing the chemical shift changes upon adding either DNA or RNA for all KH domains (Table S5, #67–72).[129]

In vitro binding assays

In vitro *transcription*

All RNA samples were *in vitro* transcribed using T7 RNA polymerase, precipitated by ethanol and purified by denaturing PAGE (12% polyacrylamide gel containing 8 M urea). The DNA templates for *in vitro* transcription are shown in Table S5 (Oligos #59–62). The gel slices were electro-eluted at 250 V in 0.5× TBE. To promote proper folding, the RNA samples were heated to 95°C for 2 min and subsequently snap-cooled on ice before use.

Fluorescent EMSA

In vitro-transcribed RNA was fluorescently labeled by ligation of pCp-Cy5 to the 3' end of the RNA with T4 RNA ligase 2. Subsequently, the reaction was purified using a spin column kit (Norgen Biotek Corp.). For binding studies, 100 nM labeled RNA in 20 mM sodium phosphate, pH 6.5, 50 mM NaCl and glycerol (15% final concentration) was incubated with increasing concentrations of $FUBP1^{N\text{-}box+KH1\text{-}4}$ (amino acids 1–457) for 15 min. Mixtures were loaded onto a 0.7% agarose gel. Gel electrophoresis was performed in 1× TBE buffer at 40 V for 4 h. Detection was performed using a Typhoon 9200 (GE Healthcare Life Sciences) at 649 nm. Data analysis was performed in Image J 2.1.0.[102] Experiments were repeated to estimate the standard deviation of the mean.

Isothermal titration calorimetry

ITC experiments were performed on a MicroCalPEAQ-ITC instrument (Malvern Panalytical) using non-isotopically labeled proteins as analyte sample and titrant or non-isotopically labeled protein as analyte and DNA oligonucleotides as titrant in NMR buffer at 25°C. U2AF2 constructs (concentration 15–30 μM) were titrated with FUBP1 N-terminal constructs (concentration 1.5–3.0 mM); FUBP1 double-KH domain constructs (concentration 20–30 μM) were titrated with DNA oligonucleotides (concentration 200–350 μM, Table S5, #64–66); *in vitro*-transcribed ssRNA (*VPS13D*, 15 μM) was titrated with $FUBP1^{KH}$ (150 μM). Binding affinity analysis was performed using MicroCalPEAQ-ITC Analysis Software (Malvern Panalytical). The standard deviations of the K_D values were estimated based on the differences in triplicate measurements.

BRET

BRET plasmid construction

The donor and acceptor vectors pcDNA3.1-cmyc-NL-GW (Addgene plasmid ID #113446), pcDNA3.1-GW-NL-cmyc (Addgene plasmid ID #113447), pcDNA3.1-GW-mCit, pcDNA3.1-mCit-GW, as well as controls pcDNA3.1-NL-cmyc (Addgene plasmid ID #113442), pcDNA3.1-PA-mCit (Addgene plasmid ID #113443), and pcDNA3.1-PA-mCit-NL-cmyc (Addgene plasmid ID #113444) were kindly provided by the Wanker group (Max-Delbrück-Centrum für Molekulare Medizin, Germany). The GATEWAY entry vectors pDON221 and pDON223 were provided by the Vidal group (Dana Farber Cancer Institute, Boston, MA). All vectors were amplified and full-length sequenced using the primers given in Table S5. Full-length wild-type ORFs being cloned into GATEWAY entry vectors were amplified from a human ORFeome collection.[130] The ORFs were full-length sequenced using primers shown in Table S5. ORFs of *FUBP1*, *SNRNP70*, and *TCERG1* (Table S6) were PCR-amplified with primers #9–10, #27–28, and #33–34, respectively (Table S5) and shuttled into pDON223 using a BP clonase II mix kit (Invitrogen). The Q5 site-directed mutagenesis kit (Invitrogen) was used to produce the following mutants: pDON223-FUBP1_A38D, pDON223-FUPB1_W586R_W615R, and pDON223-FUBP1_1-530aa (Table S6). For BRET experiments, all cDNAs were shuttled from the entry vectors into the BRET destination vectors using LR clonase technology (Invitrogen) according to the manufacturer's protocol. After the LR cloning step, the inserts were partially sequence-confirmed. All primers used are given in Table S5 and all the constructs are listed in Table S6.

Transfection

The human embryonic kidney 293 cells were transfected using Lipofectamine 2000 (Invitrogen) transfection reagent in Opti-MEM medium (Thermo Fisher) using the reverse transfection method according to the manufacturer's instructions. For BRET transfections, cells were seeded at a density of 4.0×10^4 cells per well on a white 96-well microtiter plate (Greiner) in phenol-red-free, high-glucose DMEM media (Thermo Fisher) supplemented with 5% FBS (Thermo Fisher). Transfections were performed with a total amount of 200 ng of DNA per well. If the amount of expression plasmid was less than 200 ng in a well, pcDNA3.1 (+) was used as a carrier DNA to achieve the total of 200 ng.

Experiments

Cells were transfected with plasmids encoding the acceptor (50 ng DNA) and donor (1 ng DNA). The plate was incubated for 2 days at 37°C, 5% CO_2, and 85% relative humidity prior to measurement. All measurements were performed on an Infinite M200 Pro microplate reader (Tecan). First, 100 μl of the medium was aspirated from each well. The mCitrine fluorescence was measured in intact cells (excitation/emission 513/548 nm). Then, coelenterazine h (PJK Biotech GmbH) was added at a final concentration of 5 μM. The cells were briefly shaken and incubated for 15 min inside the plate reader. After incubation, total luminescence was measured first followed by short-wavelength and long-wavelength luminescence measurements using BLUE1 (370–480 nm) and GREEN1 (520–570 nm) filters at 1,000 ms integration time. Corrected BRET (cBRET) ratios were calculated as previously described.[58] In brief, for every transfected protein pair NL-A and mCit-B, the following two control pairs were measured: NL-Stop with mCit-B and NL-A with mCit-Stop. The maximal BRET from both control pairs was subtracted from the actual test pair to correct for donor bleed-through, nonspecific binding to the tags, and background signal.

Saturation assay

For donor saturation experiments 1 ng of donor DNA encoding NL-fused proteins was co-transfected with increasing amounts of acceptor DNA encoding mCitrine-fused proteins (10, 25, 50, 100, 200, 400 ng). Fluorescence, total luminescence, and BRET were measured as described before. BRET measurements were corrected for bleed-through using NL-Stop transfections. Fluorescence and total luminescence measurements were used to estimate the amount of expressed proteins and used to plot acceptor/donor ratios on the *x*-axis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Preprocessing of RNA-seq data

Prior to genomic mapping, remaining adapter sequences were trimmed in RNA-seq data from *FUBP1* KO, *FUBP1-Nbox*^{mut}, and WT control RPE1 cells using Cutadapt v2.4.[108] A minimal overlap of 1 nt between reads and adapter was required and only reads with a length of at least 50 nt after trimming were retained for further analysis (parameters: -O 1 -m 50). Reads were mapped using STAR v2.6.1b,[107] allowing up to 4% of the mapped bases to be mismatched (--outFilterMismatchNoverLmax 0.04 --outFilterMismatchNmax 999) and a splice junction overhang (--sjdbOverhang) of 83 nt for HeLa WT samples and of 158 nt for *FUBP1* KO, *FUBP1-Nbox*^{mut}, and WT control RPE1 cells. Genome assembly and annotation of GENCODE[131] release 31 were used during mapping. Subsequently, secondary hits were removed using Samtools v1.9.[109] Exonic read counts per gene were extracted using featureCounts from the Subread tool suite v1.6.2[110] with non-default parameters --donotsort -s2.

Preprocessing of *in vivo* iCLIP data

Basic quality controls were conducted in FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) and reads were filtered based on sequencing qualities (Phred score) in the sample barcode and UMI regions using the FASTX-Toolkit v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/) and seqtk v1.3 (https://github.com/lh3/seqtk/). All reads with a Phred score below 10 in the sample barcode or UMI regions were discarded. Reads were de-multiplexed based on the sample barcode, which is found on positions 6–11 of the reads (for 6-nt sample barcodes) or on positions 4–7 (for a 4-nt sample barcode), using Flexbar v3.4.0.[111] Subsequently, barcode regions and adapter sequences were trimmed from read ends using Flexbar, requiring a minimal overlap of 1 nt of read and adapter and adding UMIs to the read identifiers. Reads shorter than 15 nt were discarded. All empty space and slash characters were removed from read identifiers in FASTQ files to prevent all information following them being lost during mapping. The downstream analysis was done as described in Chapters 3.4, 4.1, and 4.2 of Ref. [132]. Genome assembly and annotation of GENCODE[131] release 31 were used during mapping with STAR v2.6.1b.[107] The number of crosslinking events and peaks is given in Table S1. To assess the genomic distribution of iCLIP crosslink nucleotides, we used the following hierarchy: ncRNA > 3′ UTR > 5′ UTR > coding sequence (CDS) > 3′ splice site > 5′ splice site > intron > intergenic (Figure 1B). 3′ and 5′ splice site regions refer to 100 nt upstream/downstream. All other "deep-intronic" regions are called intronic regions.

Metaprofiles for *in vivo* iCLIP data

Four RNA-seq replicates from HeLa cells (imb_koenig_2018_18) served as the source for the identification of spliced introns. Mapping to the genome was performed in STAR v2.6.1b[107] (Table S1). Coordinates and number of unique supporting junction reads ("ureads") of spliced introns were extracted from the SJ file output by STAR containing high-confidence splice junctions. In the following, introns from the SJ file are called "SJ introns". SJ introns had to meet a reproducibility criterion (at least 3 out of 4 replicates). In addition, all overlapping SJ introns were removed. Finally, introns were overlaid with GENCODE release 31 annotation and filtered for level < 3, transcript support level < 4, and gene_type and transcript_type equal to "protein coding". This resulted in 88,375 SJ introns. Branch point (BP) prediction was taken from LaBranchoR.[133] LaBranchoR is based on hg19, liftOver to hg38 was done with the liftOver tool by UCSC.[134] The median distance of BP to 3′ splice sites was 25 nt. 88,008 out of 88,375 SJ introns had an annotated BP. Introns were further filtered for a minimum length of 100 nt and a maximum length of 17,000 nt. Metaprofiles were aligned at the BP. *In vivo* iCLIP replicates for each RBP were summed up and a signal threshold of 10 in the metaprofile region (−200 nt to +50 nt with respect to the BP) was imposed. Crosslinking signals per intron were normalized by "ureads" and averaged per nucleotide over all introns. For display, the normalized signal was smoothed with a Gaussian window function and window size

10. Binding enrichment for RNA maps stratified by intron length and splice site features was calculated by taking the $\log_2$ fold change of the ratio of the area under the curve (AUC) of each feature bin to the AUC in the shortest intron class or class with the weakest splice site feature. The following regions were used for AUC quantification, always with respect to BP: [−100, −25] for FUBP1, [+5, +25] for U2AF2, [−10, +10] for SF1, and [−30, −10] for SF3B1. The minimum signal in each region served as a background proxy and was taken as the lower horizontal boundary in which the AUC was calculated. For RNA maps stratified by GC content, the average GC content in the exon was contrasted to the average GC content in the first 100 nt of the downstream intron. Signal values for RNA maps aligned at 5′ splice sites were not smoothed but normalized by the average signal in the first 100 nt of the intron. RNA maps conditioned on exon rank: annotation of exons and downstream introns was extracted from GENCODE release 31. BPs were annotated as described above. SJ introns were matched to introns. Duplicated matches were resolved such that the intron with the shortest upstream exon was taken. Five exon rank classes were extracted: 1st exon, exon ranks in [2,5), [5,12), [12, 144] and second to last exon. In comparison to all other RNA maps, crosslinking signals per intron were normalized to the total crosslinking signal in the last 100 nt upstream of the 3′ splice site. "ureads" correlates with exon rank and was thus not suitable as a normalization factor. RNA maps conditioned on exon GC content: upstream exons were identified as for exon rank RNA maps. Total exon GC content over exon length was extracted. Bins are as follows: [0.07,0.41], (0.41, 0.46], (0.46, 0.53], (0.53, 0.6], (0.6, 0.91]. RNA maps condition on intron GC content: total intron GC content over intron length was extracted. Bins were as follows: [0.14, 0.36], (0.36, 0.4], (0.4, 0.46], (0.46, 0.55], (0.55, 0.9]. RNA maps for fixed intron length/differential GC content architecture followed by subsequent conditioning on differential GC content/intron length. Here, RNA binding profiles were first stratified on one class of intron length/differential GC content architecture, followed by stratification on all levels of the other factor. Binding for all RNA maps was quantified based on AUC as described above. Analyses were performed in R v4.1.1.[104]

iCLIP binding site definition (peak calling)

Binding site definition for *in vivo* iCLIP was done with PureCLIP v1.3.1. on merged replicates.[112] PureCLIP was issued with the options -iv 'chr1;chr2;chr3;' -ld -nt 4. The crosslink sites identified by PureCLIP were post-processed as previously described.[132] In detail, individual crosslink sites within a distance of 5 nt were clustered into binding regions. The binding regions were resized to obtain binding sites of a uniform width. To compare binding sites of different RBPs, we opted for 5-nt binding sites (i.e., 2 nt on either side of the position with the maximum signal) for all of the RBPs investigated (FUBP1, U2AF2, SF3B1, SF1, PTBP1). Isolated crosslink sites and binding regions of 2 nt were removed. Binding regions $\leq$ 5 nt were centered on the position with the maximum crosslink signal and extended by 2 nt on either side. Binding regions > 5 nt were divided into regions of 5 nt, by iteratively screening for the maximum signal and extending of 2 nt on either side, excluding an overlap between binding regions. Finally, at least three positions with crosslink events were required to only keep binding sites with sufficient support. To ensure sufficient support of binding sites in the individual replicates of the experiment, a reproducibility filter was applied. In order to consider the varying number and size of replicates for each experiment, we filtered for those binding sites with a total number of crosslink events higher than the 10% percentile of the distribution of crosslink counts in the single replicate. In addition, a minimum of two crosslink events was required if the 10% percentile in the replicate was below this threshold. This was required in at least two out of three, three out of four and three out of five replicates depending on the number of replicates available for the respective experiment. The numbers of called binding sites per protein are given in Table S1.

Saturation analysis

Spliced introns were identified from four RNA-seq replicates in HeLa cells (imb_koenig_2018_18) as described above. Introns were retained if they were longer than 200 nt, and if the 5′ splice site windows (the last 50 nt of the exon plus the first 75 nt of the intron) and 3′ splice site windows (the last 200 nt of the intron plus the first 20 nt of the exon) were not overlapping. 3′ splice sites overlapping to noncoding and long noncoding RNAs were excluded, resulting in 98,328 3′ splice sites. These splice sites were binned into percentiles, based on "ureads" (splice site usage) averaged over replicates. RBP binding sites were assigned to curated 3′ splice sites (the last 200 nt of the intron), requiring full overlap. For each bin, the percentage of 3′ splice sites with at least one binding site for the specific RBP was calculated (Figure 1E).

Motif enrichment for *in vivo* iCLIP

Introns were defined based on GENCODE annotation (release 31). Annotation was filtered for level < 3, transcript support level < 4, and gene_type and transcript_type equal to "protein coding", resulting in 202,623 introns. BP annotation was done as specified above. 200,199 out of 202,623 had an annotated BP. Introns were further filtered for overlaps and for having a length of at least 250 nt upstream of the defined BP. The length requirement was set to ensure that the main position of FUBP1 binding was not confounded with the 5′ splice site signal. FUBP1 binding sites (n = 854,404) were filtered for positioning within a 150-nt window upstream of the BP, resulting in 167,408 binding sites. Binding sites were ranked by their normalized signal, that is, the signal in the extended binding site (5 nt ± 5 nt) over total intron signal over intron length. Disjunct 4-mer frequencies were counted in the top/bottom 20% binding sites based on normalized signal to account for overall crosslinking preferences. Additionally, non-bound intronic regions in introns hosting the top 20% FUBP1 binding sites were also considered as an alternative background set. Here, disjunct 4-mer frequencies were calculated for all non-bound intronic regions, excluding a 20-nt region downstream of the 5′ splice site and a 150-nt region upstream of the BP. Enrichment was defined as the distance from each data point to the diagonal in a scatterplot

comparing the top 20% versus bottom 20% binding sites and, alternatively, non-bound intronic sequences. Analyses were performed in R v4.1.1.[104]

Motif enrichment upstream of branch points

Introns were extracted and BP annotated as above (200,199 introns left). Introns were further filtered for a minimum length of 500 nt and disjunct 200-nt windows upstream of the BP, resulting in 151,836 introns. Disjunct 4-mer frequencies were calculated in a position-wise manner in a 200-nt window upstream of the BP. Average background motif frequencies were calculated in a 100-nt long window 100 nt downstream of the 5′ splice site. Enrichment was defined as the distance from each data point to the diagonal in the scatterplot of position-wise frequencies versus average background frequencies.

Abundance of FUBP1 motif at 3′ splice sites

Disjunct motif occurrences were counted in a 75-nt long window 25 nt upstream of the BP. The background distribution was derived as the occurrences of nine randomly drawn motifs of length 4, repeated 100 times.

Analysis of *in vitro* iCLIP data

All samples were merged for binding site definition (peak calling) across replicates and conditions. Each *in vitro* transcript was divided into 9-nt windows, always shifted by one nucleotide. Windows were sorted by total signal and, while excluding overlapping peaks, generating a candidate. A negative binomial distribution was fit (maximum likelihood fit) to the signals on the candidate peak list. All peaks with a total signal exceeding the 90% quantile of the theoretical distribution were retained for final processing (109 peaks, see GEO record GSE220183). The background ranges were the *in vitro* transcript regions minus extended peaks (9 nt ± 5 nt). For quantifying the binding differences between conditions, replicates were averaged. Peak signals were normalized against background signals. RNA maps were based on 21 3′ splice sites present in the *in vitro* transcripts. To correct for differences in expression, nucleotide-wise signals were normalized by total *in vitro* transcript signals. Subsequently, signals were summarized per nucleotide by the 75% quantile. Replicates were averaged and subjected to Gaussian window smoothing with window size 10 before display. All analyses were performed in R v4.1.1.[104]

Analysis of oligo in vitro *iCLIP*

All data was normalized according to the total signal of all available spike-ins. Values were then extracted either per nucleotide or by binding site. Binding site positions were taken from overlays with *in vivo* U2AF2 binding sites in the intronic part of the oligonucleotide. 1,831 oligonucleotides harbored an U2AF2 binding site in the intronic part (see GEO record GSE220183). If multiple binding sites were present, that with the highest average signal in the U2AF2 samples was taken as representative. For quantifying the addition of FUBP1 on U2AF2 binding sites, only those binding sites with signal greater than the 25% quantile in one of the three replicates were considered, resulting in 1,504 binding sites. The absolute number of disjunct occurrences of the FUBP1 motif set ("TTTT" and all combinations of "TTT" and either one "A" or one "G") was counted in a 75-nt long region located 25 nt upstream of the BP. All analyses were performed in R v4.1.1.[104]

Intron length analyses of RNA-seq data

Splicing changes of *FUBP1* KO and *FUBP1-Nbox*^{mut} were analyzed with MAJIQ v2.2[135,136] with default parameter settings. MAJIQ outputs local splice variations (LSV), which were filtered as follows: for each LSV, the top two junctions in terms of absolute difference in junction usage (delta percent selected index, |ΔPSI|) were taken as representative LSVs. At least one of these two junctions needed to have an absolute ΔPSI > 0.1 and a detection probability > 0.9 (skipped for control events). Subsequently, events were filtered for exon-skipping events. Each cassette exon was then annotated with the upstream and downstream intron: genomic coordinates of the upstream/downstream intron were immediately defined in "source"/"target" events. The genomic coordinates of the respective other intron were extracted from annotation (GENCODE release 31). Overlapping cassette exons were resolved such that the event with the largest |ΔPSI| was retained (Table S3). A two-tailed Wilcoxon rank-sum test was used to assess statistical significance.

ENCODE data analysis

We retrieved raw RNA-seq data derived from an shRNA-knockdown experiment for FUBP1 in the cell line K562 from the ENCODE data portal (https://www.encodeproject.org/), using accession numbers ENCSR608IXR (*FUBP1* KD) and ENCSR260BQC (control). Alignment was performed in STAR (version 2.7.8a)[107] with standard ENCODE options. We applied MAJIQ v2.3[135,136] to identify and quantify cassette exons in the RNA-seq data. First, a splice graph was built on the BAM files and the GENCODE gene annotation (v38, human genome version hg38). Then, the difference in junction usage between knockdown and control samples was calculated (as ΔPSI). Next, alternative splicing events such as cassette exons (CEs) were categorized and quantified in the splicing graph using MAJIQ Modulizer. Probabilities were calculated for each junction, testing for $|\Delta PSI| > 0.05$ (probability changing [P_s]) and $|\Delta PSI| < 0.02$ (probability non-changing [P_n]). The MAJIQ Modulizer output was then processed in R, filtering for significantly regulated CEs and a control group with unregulated CEs. A CE is defined as significantly regulated if $|\Delta PSI| \geq 0.055$ for all junctions, $P_s \geq 0.9$ for at least one junction pair (inclusion junction + skipping junction), the sign within both junction pairs is inverse, and within the junction pairs the lower |ΔPSI| is at least 50% of the higher |ΔPSI|. A CE is considered to be unregulated if $P_n \geq 0.5$ and $|\Delta PSI| \leq 0.02$ for all junctions. Overall, this resulted in a total of 173 significantly regulated CEs and a control group with 1,910 unregulated CEs for further

analysis. To categorize CEs into more included and less included, a representative ΔPSI was chosen for each CE based on the maximum |ΔPSI| of both inclusion junctions. Based on this, there were 30 more-included and 143 less-included exons.

Splicing changes upon FUBP1 LoF mutations

Significant differentially spliced exon-skipping events upon (i) loss-of-function (LoF) mutations of *FUBP1* in low-grade gliomas (37 events), (ii) in *FUBP1* siRNA knockdown in U87MG cells (109 events) and (iii) LoF mutations of other splicing factors (433) were extracted from Seiler et al.[1] Junction lengths comprise the upstream intron, the skipped exon and the downstream intron. A two-tailed Wilcoxon rank sum test was used to assess statistical significance.

Mutations in FUBP1 in cancer patients

We searched multiple databases to identify disease-related mutations within the *FUBP1* gene. We focused on the minimal binding interface to U2AF2 (FUBP1 amino acids 25–56) to find mutations that potentially abolish the interaction with the U2AF2 RRM2 domain. The following databases were used: ICGC Data Portal,[137] cBioPortal,[138,139] Exac,[140] Cosmic,[141] GDC Data Portal,[142] gnomAD,[140] and ClinVar.[143] All cancer-related mutations in *FUBP1* in the observed region and the underlying cancer type are listed in Figure S4B.

Scoring of splice site features

3′ and 5′ splice site strength was scored with MaxEnt scan.[144] Py tract strength was determined as follows: a 39-nt region upstream of the AG dinucleotide at the 3' splice site was screened with sliding windows of increasing length (width 5–30 nt) to identify the window with the highest Py tract strength. The Py tract strength of each window was calculated as the X^2 test statistic with 1 degree of freedom, comparing the observed number of pyrimidines with the expected number based on the assumption of a uniform nucleotide distribution. In addition, candidate Py tracts were required to end within 10 nt upstream of the AG dinucleotide. Using this approach, the median length of identified Py tracts was 16 nt. BP strength was assessed according to the U2 binding energy, that is, the number of hydrogen bonds between the candidate sequences and the BP binding sequences in the U2 snRNA. Hydrogen bonds form between A:T (2 bonds), G:C (2 bonds), and G:U (1 bond; in fact also 2 bonds, but punished for being a wobble base pair) with the BP nucleotide bulging out and being omitted from the pairings. The Vienna RNA package v2.4.17[113] (RNAduplex) was used to determine the optimal hybridization structure between U2 snRNA sequences (GUGUAGUA) and the motif (position −5 to +3, excluding the BP nucleotide). Predicted binding energy was the determined sum of hydrogen bonds forming between complementary motifs and U2 snRNA nucleotides.

Evolutionary analyses

We annotated the domain architecture of FUBP1 using the function annoFAS provided in the FAS package[106] (https://github.com/BIONF/FAS). The domain architecture-aware phylogenetic profile of FUBP1 across 174 mammals, 274 non-mammalian vertebrates, 277 invertebrates, 410 fungal species, 94 protozoa, and 145 plants was generated with the targeted ortholog search tool fDOG (https://github.com/BIONF/fDOG)[145] using the human FUBP1 (UniProt: Q96AE4) as a seed. fDOG was run with the options --minDist class, --maxDist phylum, –checkCoorthologsRef, and --countercheck. *Homo sapiens* (GenBank: GCF000001405) served as the reference taxon. Intron length and GC content information was extracted based on the respective gff and fasta files downloaded from NCBI RefSeq Genome. Intron length estimates and motif searches were performed in R v4.0.5. A/B box presence in the human proteome was determined as follows: in brief, we used the shell command *grep* to search for the regular expression "[ST][AK][QA]W..YY[RK]" in 19,519 human proteins encoded in the NCBI RefSeq Genome assembly GCF_000001405.39. The resulting three hits were NCBI: XP_011540693.1 (FUBP1, 2 motif instances), NCBI: NP_003925.1 (FUBP3, 1 motif instance), and NCBI: NP_001353228.1 (KHSRP, 3 motif instances). For counting FUBP1 motif occurrences across species, intron definitions were extracted for all the species investigated and motifs were counted in a 25-nt window located 25 nt upstream of the 3′ splice site.

Analysis of RBP crosslinking to snRNAs

In vivo iCLIP data from FUBP1, U2AF, SF1, SF3B1, and PTB was remapped to a custom database consisting of snRNAs, tRNAs, and rRNAs using STAR v2.7.3a.[107] Specifically, RNU1-1, RNU2-1, RNU4-1, RNU6-1, RNU5D-1, RNU7-1, RNU11, RNU12, RNU4ATAC, and RNU6ATAC were included. tRNA coordinates were retrieved from GtRNAdb (data release 19). "hg38-tRNAs.fasta", containing 429 high-confidence tRNA annotations, was downloaded. Because tRNAs are quite similar when stratified on their carried amino acid, one representative tRNA was selected per amino acid (tRNA with "1-1" in the name). In summary, this resulted in 22 tRNAs. Finally, the following rRNAs were added: 12S_gi, 16S_gi, 18S_gi, 28S_gi, 5.8S_gi, and 5S_gi. Mapping steps were performed as follows: all sequences were furnished with one additional base upstream of the sequence with the rationale of being able to display iCLIP coverage of reads starting directly at the 5' end of the sequence. tRNAs and snRNAs were furnished with the actual base upstream of the sequence. rRNAs were furnished with an "N". Reads were mapped per replicate with STAR v2.7.3a using the settings described above for *in vivo* iCLIP samples. Few reads were mapped to the minus strand and thus removed. Uniquely mapping reads were subjected to duplicate removal based on identical UMIs (--method unique) using UMI-tools v1.0.0.[146] Based on the remaining reads, iCLIP coverage profiles were exported as well as count tables containing the number of reads overlapping the genomic ranges of the defined RNAs.

Subnuclear distribution of FUBP1-bound genes

The subnuclear spatial distribution for introns in HeLa cells was taken from Tammer et al.[64], in which Chrom3D, a 3D genome-modeling tool that integrates 3DHi-C data and ChIP-seq data was used to assign distances from the nuclear center for topologically associated domains. The distance from the nuclear center is described by five concentric radial scopes where 1-to-5 point to the center–periphery axis. Our *in vivo* iCLIP data from SF3B1, FUBP1, and U2AF2 was then overlaid with the reported introns and the percentage of bound introns was counted. Enrichment was calculated as the percentage of bound introns in each radial scope compared to the first.

Mathematical modeling

Topology of the exon definition model

Splicing reactions are catalyzed by the spliceosome, which recognizes splice site sequences and forms a catalytically active higher-order complex across introns. To model this process, we considered that human spliceosomes frequently operate by a so-called "exon definition" mechanism, in which the pioneering spliceosome subunits U1 and U2 cooperatively bind to splice sites flanking an exon before the final cross-intron complex is formed during spliceosome maturation.[86] Because the initial binding of U1 and U2 plays a decisive role in splicing decisions,[86] we model only the initial exon definition step and assume the corresponding binding patterns determine splicing outcomes, as described below.

In the model pre-mRNA, none of the three exons are bound ("defined") by the spliceosome (white boxes), therefore this state is denoted "P0_0_0" (Figure S7F) with the notation "_" indicating the presence of an intron. In the model, the pre-mRNA (P0_0_0) is synthesized at a constant rate s. The spliceosome can bind reversibly to each of the exons with on-rates k_1, k_2, and k_3. For instance, from P0_0_0 we can obtain P1_0_0, P0_1_0, and P0_0_1 through binding to the first, second, and third exon, respectively. Subsequent binding is possible; for example, P1_0_1 can be generated from P1_0_0 with the rate constant k_3. In total, there are eight spliceosomal binding states, including the fully bound state (P1_1_1), in which all exons are defined. All binding reactions are assumed to be reversible, i.e., k_4, k_5, and k_6 are the dissociation rate constants and the reverse of k_1, k_2, and k_3, respectively. For example, in state P1_1_0, spliceosome dissociation from exon 1 with the rate constant k_4 yields the species P0_1_0.

Depending on the exon definition states, splicing decisions are made, and irreversible splicing reactions are possible. For a splicing event to occur, we consider that both exons flanking a future splice junction must be defined. For instance, skipping of exon 2 is possible from P1_0_1 and occurs with the rate constant i_{12}. Likewise, splicing of the first intron occurs from the species P1_1_0 and P1_1_1 (rate constant i_1), and splicing of the second intron from P0_1_1 and P1_1_1 (rate constant i_2). The inclusion isoform is generated in two steps, that is, from the subsequent removal of introns 1 and 2 in random order: from the binding state P1_1_1, intron splicing generates two alternative intermediates in which either of the introns is already spliced (P1_11 or P11_1) and the retained intron can be further spliced in a subsequent reaction. Splicing of the partially defined species P1_1_0 and P0_1_1 yields the species P11_0 and P0_11; in these, the spliceosome can further reversibly bind exons 3 and 1, respectively, and undergo a second splicing reaction toward inclusion. In the model, all terminal splice products are subject to degradation (k_{incl}, degradation rate constant of inclusion; k_{skip}, skipping; k_{dr1}, first intron retention; k_{dr2}, second intron retention). The degradation rate constant of the full intron retention isoform is the sum of k_{dr1} and k_{dr2}, reflecting that either intron may contain a destabilizing premature stop codon. Model species that can be bound or spliced further (P0_0_0, P1_0_0, P0_1_0, P0_0_1, P1_1_0, P1_0_1, P0_1_1, P1_1_1, P0_11, P1_11, P11_0, P11_1) are not subject to degradation, but they can be exported from the nucleus with the rate constant k_{ret}. This reaction reflects that there is a limited time window for splicing to occur, the intermediates otherwise being terminally frozen in the corresponding intron retention state. The ordinary differential equations of the model are given in Table S4.

Topology of the intron definition model

Because a subset of human genes are spliced by an intron definition mechanism, we also considered this scenario in a modified version of our splicing model. In contrast to the exon definition model, the 5′ and 3′ splice sites of an exon can be bound independently of one another in the intron definition model. Furthermore, splicing of an intron is possible as soon as both splice sites flanking this intron are defined. Hence, definition of two splice sites is sufficient for splicing to occur, whereas in the exon definition model four splice sites need to be defined (3′ and 5′ splice sites of the two flanking exons). For the intron definition model, we use a notation for binding state similar to that for exon definition. For instance, for consistency, we assigned the state in which no spliceosome component is bound as P0_0_0. For spliceosome binding to exons 1 and 3, we again considered a single binding reaction, as only the splice sites flanking the intron of interest are relevant for splicing. Hence, a transition from "0" to "1" in the first position (e.g., P0_0_0 to P1_0_0) represents a spliceosome binding state downstream of exon 1 (5' of the first intron), and "0" to "1" in the third position indicates binding upstream of exon 3 (3' of the second intron). For exon 2, we treat splice-site binding as two separate events. We use "0" to denote no binding, "a" for upstream binding (e.g., P0_a_0), "b" for downstream binding (e.g., P0_b_0), and "1" for both U2 and U1 being simultaneously bound (e.g., P0_1_0). Again, the presence or absence of "_" indicates whether or not the intron is removed. We adopted the same parameter notation, that is, k_1/k_4 and k_3/k_6 to describe binding/dissociation at exons 1 and 3, respectively. The new parameters k_{2a}/k_{5a} (upstream) and k_{2b}/k_{5b} (downstream) were introduced to represent spliceosome binding/dissociation around exon 2. There are a total of 16 spliceosomal binding states in the intron definition model, with the following additional states not part of the exon definition model: P0_a_0, P0_b_0, P1_a_0, P1_b_0, P0_a_1, P0_b_1, P1_a_1, and P1_b_1. If both splice sites flanking a future splice junction are defined, splicing decisions, implemented as irreversible splicing reactions in the model, can occur. Skipping of exon 2 is possible from P1_0_1 and occurs with the rate i_{12}. Splicing of the first intron occurs from species P1_a_0, P1_1_0,

P1_a_1, and P1_1_1 (rate i_1), and splicing of the second intron occurs from P0_b_1, P0_1_1, P1_b_1, and P1_1_1 (rate i_2). The inclusion isoform is generated in two steps: first, intron 1 or 2 is spliced from P1_1_1, generating P1_11 or P11_1, respectively. Second, the retained intron can be further spliced in a subsequent reaction. Splicing of the partially defined species P1_a_0, P1_1_0, P0_b_1, and P0_1_1 yields the species P1a_0, P11_0, P0_b1, and P0_11, respectively. To these, the spliceosome can bind further reversibly with the association rate constants k_1, k_{2a}, k_2, and k_3 (depending on the site of binding), and if the species P1_11 or P11_1 are formed, a second splicing reaction toward inclusion can occur. All terminal splice products are subject to degradation, for which we adopted the same assumptions and notation as for the exon definition model. Again, model species that can be bound or spliced further (P0_0_0, P1_0_0, P0_a_0, P0_b_0, P0_1_0, P0_0_1, P1_1_0, P1_a_0, P1_b_0, P1_0_1, P0_a_1, P0_b_1, P0_1_1, P1_a_1, P1_b_1, P1_1_1) can be exported from the nucleus with the rate constant k_{ret}. The ordinary differential equations of the model are given in Table S4.

Model simulation and analysis

The differential equations were implemented in Matlab 2020b and solved using ode15s. To analyze splicing outcomes, we assumed a steady state, and performed numerical simulations over long time periods (t = 1,000,000 min) to ensure that the concentrations of the model species remained constant. Thus, we consider an RNA sequencing experiment, in which gene expression was measured in a stationary cell population in the absence of any external perturbation. As a measure of splicing outcome, we used the steady-state concentrations of inclusion and skipping (see also below).

Genome-wide splicing modeling by parameter sampling

The exon definition model consists of 15 kinetic parameters which belong to the following classes of reactions: spliceosome binding (k_1, k_{2a}, k_{2b}, k_3), spliceosome dissociation (k_4, k_{5a}, k_{5b}, k_6), splicing catalysis (i_1, i_2, i_{12}), and others, which are rates of pre-mRNA synthesis (s), mRNA degradation (k_{int}, k_{skip}, k_{dr1}, k_{dr2}), and terminal intron retention (k_{ret}). The values of these parameters were unknown and likely greatly differ between exons in the human genome. To mimic the heterogeneity of exons in the human genome and to assess the robustness of our simulation results, we randomly sampled all kinetic parameters in our model 10,000 times. As a reference parameter set, all parameter values were set to 1, except for k_{ret}, k_{incl}, and k_{skip}, which were set to 0.01 to ensure low levels of intron retention that are typically observed in RNA sequencing datasets. We sampled each parameter in the model within a +/-sevenfold range around this reference using Latin hypercube sampling (lhsdesign command in Matlab). We performed simulations for each parameter realization and calculated PSI = inclusion / (inclusion + skipping) as a measure of alternative splicing. We obtained a PSI distribution between 0 and 1 that closely resembled the experimentally measured genome-wide PSI in control cells. The same procedure was applied for intron definition, with the only difference being the number of parameters involved -17 in this case. These kinetic parameters belong to the following classes of reactions: spliceosome binding (k_1, k_{2a}, k_{2b}, k_3) and spliceosome dissociation (k_4, k_{5a}, k_{5b}, k_6); the remainder are identical to those used for exon definition.

Modeling FUBP1 knockout effects

To reproduce the *FUBP1* KO data, we implemented two distinct assumptions about the mechanism of action of FUBP1: that FUBP1 affects late spliceosomal catalysis (i.e., the rate constants i_1, i_2 and/or i_{12}), or that FUBP1 affects early spliceosomal binding (i.e., the rate constants k_1–k_6). For both mechanistic assumptions, we considered that FUBP1 predominantly binds long introns (Figure 6A). When simulating the effect *FUBP1* KO has on splicing catalysis (model 2 in Figure 6A), we assumed that the splicing of short introns is unaffected, but that KO selectively reduces the splicing rate for the excision of long introns 3.5-fold compared to control. To reflect different combinations of long and short introns, we considered three scenarios in the *FUBP1* KO simulations: (i) for the simulation of cassette exons flanked by two long introns, we assumed that the *FUBP1* KO slows all three splicing reactions in the model, that is, the excision of intron 1, excision of intron 2 and exon skipping (i_1, i_2, and i_{12} are changed). (ii) For exons flanked by one short and one long intron, it was assumed that the splicing rate of the short intron is unaffected by *FUBP1* KO, whereas splicing rates of the long intron and skipping are reduced. The long intron was either considered to be located upstream of the alternative exon (ii.a: i_1 and i_{12} are changed) or downstream (ii.b: i_2 and i_{12} are changed). In either case, the skipping reaction was considered as an FUBP1-dependent, long-range splicing event and was therefore perturbed in the *FUBP1* KO simulation (i_{12} is changed). (iii) The third hypothetical scenario, in which an alternative exon is flanked by two short introns, was not explicitly considered in our simulations, as the model would predict no PSI change upon *FUBP1* KO in this case. For each parameter sample (hypothetical exon), the KO scenarios i, ii.a, and ii.b were implemented separately, resulting in three sets of 10,000 KO simulations. For each of these, the PSI changes upon *FUBP1* KO were calculated [ΔPSI = PSI(KO) – PSI(control)], and the corresponding ΔPSI distribution (Figure 6B) agrees well with the experimental observation in RNA sequencing experiments. In the alternative *FUBP1* KO implementation (model 1 in Figure 6A), we assumed that FUBP1 promotes initial U2 binding to the 3′ splice site. Because the 3′ splice site marks the downstream end of an intron, we assume that the *FUBP1* KO reduces spliceosome binding to exons located downstream of long introns. In our model, a long intron 1, therefore, results in a reduced exon 2 definition rate upon *FUBP1* KO (k_2 changed 1.7-fold compared to control). Likewise, a long intron 2 diminishes exon 3 definition (k_3 changed 1.7-fold upon *FUBP1* KO). These perturbations were implemented alone (one long and one short intron), or in combination (two long introns), and the corresponding ΔPSI distributions across all 10,000 parameter realizations are shown in Figure 6B. The perturbation in binding parameters (k_2, k_3) was chosen to be smaller (1.7-fold) than the effect on splicing parameters (3.5-fold, model described above) to adjust for similar-sized effects on splicing in both implementations. In contrast to the *FUBP1* KO RNA sequencing data, these spliceosome binding simulations predict opposite PSI changes for short introns being located upstream or downstream of the alternative exon. Hence, a model in which FUBP1 enhances the catalytic excision of long introns explains the *FUBP1* KO data better when compared to a model in which FUBP1 primarily helps to

recruit the pioneering U2 subunit to the 3′ splice site. The same *FUBP1* KO simulations were also implemented in the intron definition scenario. Here, the effect of FUBP1 on spliceosome binding (model 1 in Figure 6A) was assumed to affect the k_{2a} parameter for a long upstream intron and k_3 for long downstream introns. If both introns are long, FUBP1 influences both k_{2a} and k_3. The effect of FUBP1 on splicing catalysis (model 2 in Figure 6A) in the intron definition model was implemented in the same way as described above for the exon definition model. For FUBP1-based mechanisms of action, that is, binding and catalysis effects, very similar results were observed for the intron and exon definition scenarios (Figure S7G). Hence, the model's prediction that FUBP1 affects splicing catalysis is robust and does not depend on the mechanism of splicing decision making.