The role of FUBP1 in splicing, disease and evolution

Dissertation zur Erlangung des Grades

“Doktor der Naturwissenschaft“

Am Fachbereich Biologie
der Johannes Gutenberg Universität Mainz

Miriam Magdalene Mulorz
geb. am 02.11.1992 in Speyer

Mainz, Oktober 2023

Dekan: Prof. Dr. Eckhard Thines

1. Berichterstatter: [redacted]

2. Berichterstatter: [redacted]

Tag der mündlichen Prüfung: 20.02.2024

Affidavit

I wrote this PhD Thesis without the help of a third party. All sources and aids have been described. For the publications in chapter 3 and 4, I stated my contribution and role in the respective studies. In chapter 5, I designed and carried out all described experiments. For the nascent RNA-seq, library preparation and sequencing were performed by the [redacted], preprocessing was done by [redacted] and data analysis was performed by [redacted]

Das Wichtigste ist, das man nicht aufhört zu fragen.

– Albert Einstein

Summary

Splicing is the process of intron excision and exon re-ligation, and it is a crucial step of gene regulation. Mechanistically, splicing is well studied. However, splicing is a highly regulated process and there are multitudes of *cis*-regulatory and *trans*-acting elements that control splicing. This so-called "splicing code" is less understood, and in this study, we employed an array of high-throughput sequencing methods, coupled to structural biology, mathematical modeling and molecular biological experiments, to examine the role of the Far-Upstream Binding Protein FUBP1 in splicing and further, the regulation of *CD19* mRNA splicing.

Previous studies showed that the transcription factor FUBP1 also regulates certain splicing events. Additionally, [redacted] demonstrated that FUBP1 stabilizes U2AF2, a core splicing component, to the py-tract. So far, however, the transcriptome-wide role of FUBP1 in splicing remains elusive. We showed that FUBP1 binds to a hitherto unknown U-rich sequence upstream of the branch point. It binds more than 80% of all pre-mRNA, rendering it a core splicing factor. Using NMR-based technologies and reporter assays, we showed that FUBP1 interacts with proteins of both the 3'ss and the 5'ss. We found that the N-terminal N-box of FUBP1 interacts with the RRM2 domain of U2AF2. In addition, an animal-specific C-terminal A/B-box of FUBP1 interacts with proline-rich stretches of SF1. RNA-seq of two FUBP1 mutant cell lines generated by CRISPR/Cas9 revealed that exons flanked by long introns are skipped when FUBP1 is lost. Analyzing RNA-seq of low-grade glioma patients with FUBP1 deficiencies yielded comparable results, indicating that FUBP1 facilitates splicing of long introns. We hypothesize that FUBP1 evolved to regulate splicing of long introns in higher eukaryotes by interacting with both the 3'ss and the 5'ss, connecting the splice sites.

In a second study, we dissected the elements that are important for splicing of *CD19* mRNA. B cell acute lymphoblastic leukemia (B-ALL) patients can be treated with the cell-based immunotherapy CART-19, which specifically targets CD19-expressing cells. However, they often experience relapse due to CD19 epitope loss. Using a high-throughput mutagenesis approach, we detected over 200 mutations that lead to approx. 100 cryptic isoforms that can contribute to epitope loss. By cloning a selection of the mutations into a *CD19* minigene and performing RT-PCR, we could confirm these results. In addition, shRNA-induced KD of splicing factors revealed that *trans*-acting factors like SF3B4 and PTBP1 are essential for correct *CD19* splicing. With this study, we offer a comprehensive overview of the *CD19* splicing code.

Taken together, this PhD thesis highlights the interdependency between *cis*- and *trans*-acting splicing components. It underlines the importance of splicing – and the knowledge thereof – in the context of health and disease.

Zusammenfassung

Spleißen ist der Prozess der Entfernung von Introns und dem Verbinden von Exons. Es ist ein zentraler Teil der Genregulation. Mechanistisch ist Spleißen gut untersucht. Aber Spleißen ist ein stark regulierter Prozess mit einer Vielzahl von *cis*-regulierenden und *trans*-agierenden Elementen, die das Spleißen beeinflussen. Dieser sogenannte „Spleißschlüssel" ist weniger gut verstanden. In dieser Studie verwendeten wir mehrere Hochdurchsatzsequenziermethoden zusammen mit Strukturbiologie, mathematischem Modellieren und Molekularbiologie, um den Einfluss von FUBP1 auf das Spleißen zu erforschen und um die Spleißregulation von *CD19* mRNA zu verstehen.

Studien fanden heraus, dass der Transkriptionsfaktor FUBP1 auch bestimmte Spleißereignisse reguliert. Unsere Arbeitsgruppe zeigte, dass FUBP1 U2AF2, eine zentrale Spleißkomponente, am Py-Trakt stabilisiert. Die transkriptomweite Rolle von FUBP1 während des Spleißens bleibt jedoch unerforscht. Wir zeigten hier, dass FUBP1 an eine bisher unbekannte, U-reiche Sequenz oberhalb des Verzweigungspunkt bindet. Es bindet mehr als 80% aller prä-mRNA, was es zu einem zentralen Spleißfaktor macht. Mit NMR-basierten Technologien und Reporterassays konnten wir zeigen, dass FUBP1 sowohl mit Proteinen der 3' als auch der 5' Spleißstelle interagiert. Wir stellten fest, dass die N-terminale N-box von FUBP1 mit der RRM2-Domäne von U2AF2 interagiert. RNA-Sequenzierung von zwei FUBP1-Mutatenzelllinien, die mit CRISPR/Cas9 generiert wurden, ergab, dass Exons, die von langen Introns umgeben sind, in der Abwesenheit von FUBP1 übersprungen werden. Analysen von RNA-Sequenzierungen von Patienten mit niedriggradigen Gliomen und FUBP1-Defizienz ergab vergleichbare Ergebnisse, was darauf hinweist, dass FUBP1 das Spleißen langer Introns ermöglicht. Wir vermuten, dass FUBP1 sich evolutionär entwickelte, um das Spleißen langer Introns in höheren Eukaryoten zu regulieren, indem es die 3' und 5' Spleißstellen verbindet.

In einer zweiten Studie analysierten wir Elemente, die für das Spleißen der *CD19* mRNA relevant sind. Patienten mit akuter lymphoblastischer B-Zellleukämie (B-ALL) können mit der zellbasierten CART-19 Immuntherapie behandelt werden, die spezifisch CD19-exprimierende Zellen eliminiert. Dennoch können Patienten einen Rückfall durch den Verlust des CD19-Epitops erleiden. Mithilfe einer Hochdurchsatzmutagenese konnten wir 200 Mutationen detektieren, die zu ca. 100 kryptischen Isoformen und damit zu Epitopverlust führen können. Wir klonierten eine Auswahl dieser Mutationen in ein *CD19* Minigen und führten eine RT-PCR durch, die diese Ergebnisse bestätigte. Außerdem konnte eine shRNA-induzierte Reduktion verschiedener Spleißfaktoren zeigen, dass *trans*-agierende Faktoren wie SF3B4 und PTBP1 für das Spleißen von *CD19* essenziell sind. Mit dieser Studie bieten wir einen umfassenden Überblick über den *CD19* Spleißschlüssel. Zusammengefasst beleuchtet diese Doktorarbeit die Abhängigkeit von *cis*- und *trans*-agierenden Spleißkomponenten. Es unterstreicht die Wichtigkeit von Spleißen – und das Wissen darüber – in Gesundheit und Krankheit.

Table of Content

1. Introduction

1.1. Introduction into splicing

1.1.1. Discovery of splicing

When the groups of Philip A. Sharp and Richard J. Roberts published the landmark discovery of Splicing in 1977 (S. M. Berget, Moore, and Sharp 1977; Chow et al. 1977), physicist and biochemist Walter Gilbert described it in following words: “Our picture of the organisation of genes in higher organisms has recently undergone a revolution. Analyses of eukaryotic genes […] suggest that in general coding sequences on DNA, the regions that will ultimately be translated into amino acid sequence, are not continuous but are interrupted by 'silent' DNA” (Gilbert 1978). This view was also shared by the Nobel Prize Committee, which awarded Sharp and Richard J. Roberts with the Nobel Prize in Medicine (The Nobel Assembly at Karolinska Institute 1993). Only three years later, the first mammalian splicing event was discovered, when it became apparent that both membrane-bound and secreted IgM immunoglobulins, which have two distinct mRNA isoforms, are transcribed from the same gene (Alt et al. 1980; Early et al. 1980; Nilsen and Graveley 2010) Since then, our understanding of splicing has increased tremendously.

RNA splicing, as we know today, is an essential layer of gene regulation in all eukaryotes (Ben-Dov et al. 2008; Blencowe 2006; Klenerman, Cerundolo, and Dunbar 2002; Matlin, Clark, and Smith 2005). It occurs in >95% of all human genes and is essential to form the correct messenger RNA (mRNA) from pre-mRNA (Baralle and Giudice 2017; Gallego-Paez et al. 2017). During splicing, the “silent DNA”, called introns, is excised and the remaining protein-coding introns are re-ligated (S. M. Berget, Moore, and Sharp 1977; Wahl, Will, and Lührmann 2009). The spliceosome, a megadalton ribonucleoprotein (RNP) complex, catalyzes the splicing reaction (Wilkinson, Charenton, and Nagai 2020). For an intron to be recognized and spliced, three *cis*-regulatory sequences are needed, namely the 5’ and the 3’ splice site (5’ss and 3’ss), which mark the beginning and end of an intron, respectively. Further, the conserved branch point adenosine is important for the intron to form a lariat. In addition to less conserved *cis*-regulatory elements like the polypyrimidine tract downstream of the branch point, there are more than 150 *trans*-acting factors that regulate splicing (Gerstberger, Hafner, and Tuschl 2014; Wilkinson, Charenton, and Nagai 2020).

Despite tremendous advances in splicing research, many aspects of this intricate process remain to be unraveled. The involvement of various RNA motifs and RNA-binding proteins emphasizes the critical role of splicing. Therefore, further study of splicing is warranted to deepen our understanding of this fundamental biological process.

1.1.2. Alternative splicing

Although splicing constitutes a huge undertaking for any cell and organism, it also provides substantial gene regulatory opportunities. The exon-intron architecture allows for the phenomenon of alternative splicing, a mechanism that endows cells with the ability to differentially include or exclude both exons and introns. In humans, an estimated 95-100% of human genes can be alternatively spliced. Alternative exons can undergo exon inclusion or exon skipping (ES), while the lack of intron removal is called intron retention (IR). Subtypes of exon skipping are usage of alternative 3'ss or 5'ss (A3'SS, A5'SS), or mutually exclusive exons (Nilsen and Graveley 2010; Pan et al. 2008; E. T. Wang et al. 2008) (**Figure 1**). Alternative splicing harbors the potential for evolutionary and proteomic diversity resulting in an increase by fourfold from genes to proteins in humans (Keren, Lev-Maor, and Ast 2010; Modrek and Lee 2002). This diversity is pivotal for cell identity or regulatory purposes.

For example, in healthy pancreatic beta cells, the relative expression level of the splicing enhancer SRSF1 regulates skipping and inclusion of the alternative exon 11 of the insulin receptor mRNA *ISNR*, which results in INSR-A or ISNR-B, respectively. This plays a crucial role in glucose level regulation within the cell (Dlamini, Mokoena, and Hull 2017; Malakar et al. 2016). Perhaps the most popular example illuminating of the potential of alternative splicing is the *D. melanogaster* gene *Down syndrome cell adhesion molecule* (*Dscam*), which astonishingly generates over 38,000 distinct isoforms, in an organism containing only 14,500 genes (Schmucker et al. 2000).

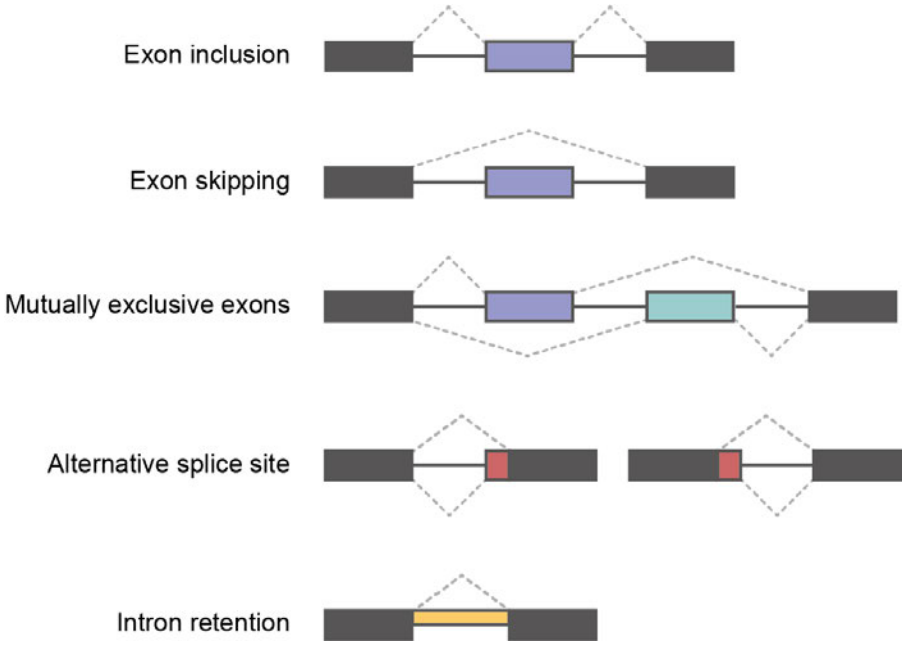


Figure 1: Modes of alternative splicing. The major modes of alternative splicing are exon skipping, mutually exclusive exons, and intron retention. Alternative splice site usage is considered a variant of exon skipping.

This directly raises the question of why introns are needed in general and why it is not possible to reorganize exons without transcribing and excising the majority of a transcript, imposing an enormous energetic burden to the organism. Intriguingly, introns seem to profoundly influence the mRNA life cycle (Jo and Choi 2015). Most directly, introns can affect splicing decisions, as alternative exons surrounded by long introns are prone to being skipped (Pan et al. 2008; Roy et al. 2008; Sorek and Ast 2003). In addition, introns can harbor regulatory sites that either enhance or inhibit inclusion of a target exon (Maniatis and Tasic 2002). Surprisingly, introns can also partake in gene expression, mRNA export and capping of its transcript (Jo and Choi 2015). Furthermore, introns can serve as a source of *de novo* exons. Those exonization events arise when mutations or gene editing introduce new splice sites (Keren, Lev-Maor, and Ast 2010). After splicing, introns are rapidly degraded, but intriguingly, it has been shown that in yeast, several introns exert a higher level of stability and persist in the cell to promote cell growth (Morgan, Fink, and Bartel 2019).

1.2. Molecular mechanisms of splicing

1.2.1. Biochemistry of splicing in humans

Two types of spliceosome exist: the major spliceosome, or U2-dependent spliceosome; and the minor spliceosome, or U12-dependent spliceosome. This thesis centers on the major spliceosome, which uses canonical splice sites and is responsible for more than 99% of all splicing events (Bergsma et al. 2018; Turunen et al. 2013). Across all eukaryotes, the three most important *cis*-elements within the RNA are conserved. The 5' splice site (5'ss) starts with a "GU" sequence in metazoans and marks the beginning of the intron, and the 3'splice site (3'ss), the "AG"-sequence defines the end of the intron. Roughly 20-50 nt upstream of the 3'ss, the branch point adenosine can be found within a degenerated branch point sequence in humans (Gao et al. 2008). During splicing, the branch point attacks the 5'ss, forming a lariat structure. The unbound end of the 5'ss then attacks the end of the 3'ss, ligating the exons together and releasing the lariat. This two-step esterification reaction is catalyzed by the spliceosome (Wilkinson, Charenton, and Nagai 2020).

The human major spliceosome comprises five snRNPs: U1 and U2 snRNP, along with the U4/U6.U5 tri-snRNP. These complexes interact in a well-coordinated, sequential manner with *cis*-elements of pre-mRNA. Each snRNP includes a U-rich small nuclear RNA (snRNA) encircled by seven homologous Sm proteins, except for U6 snRNP, which associates with LSm proteins. Additional snRNP-specific proteins further contribute to the splicing process. To start splicing, first, the intron-exon boundaries must be defined. The U1 snRNA of U1 snRNP base pairs with the 5'ss, and U2 auxiliary factors 1 and 2 (U2AF1 and U2AF2) bind the 3'ss and polypyrimidine tract (py-tract), respectively. Together with the branch point binding protein SF1, those factors form the complex E, marking the initial splicing step (Wilkinson, Charenton, and Nagai 2020). With this, the definition of the intron is concluded, but how they gain specificity, is not completely understood (**Figure 2**A).

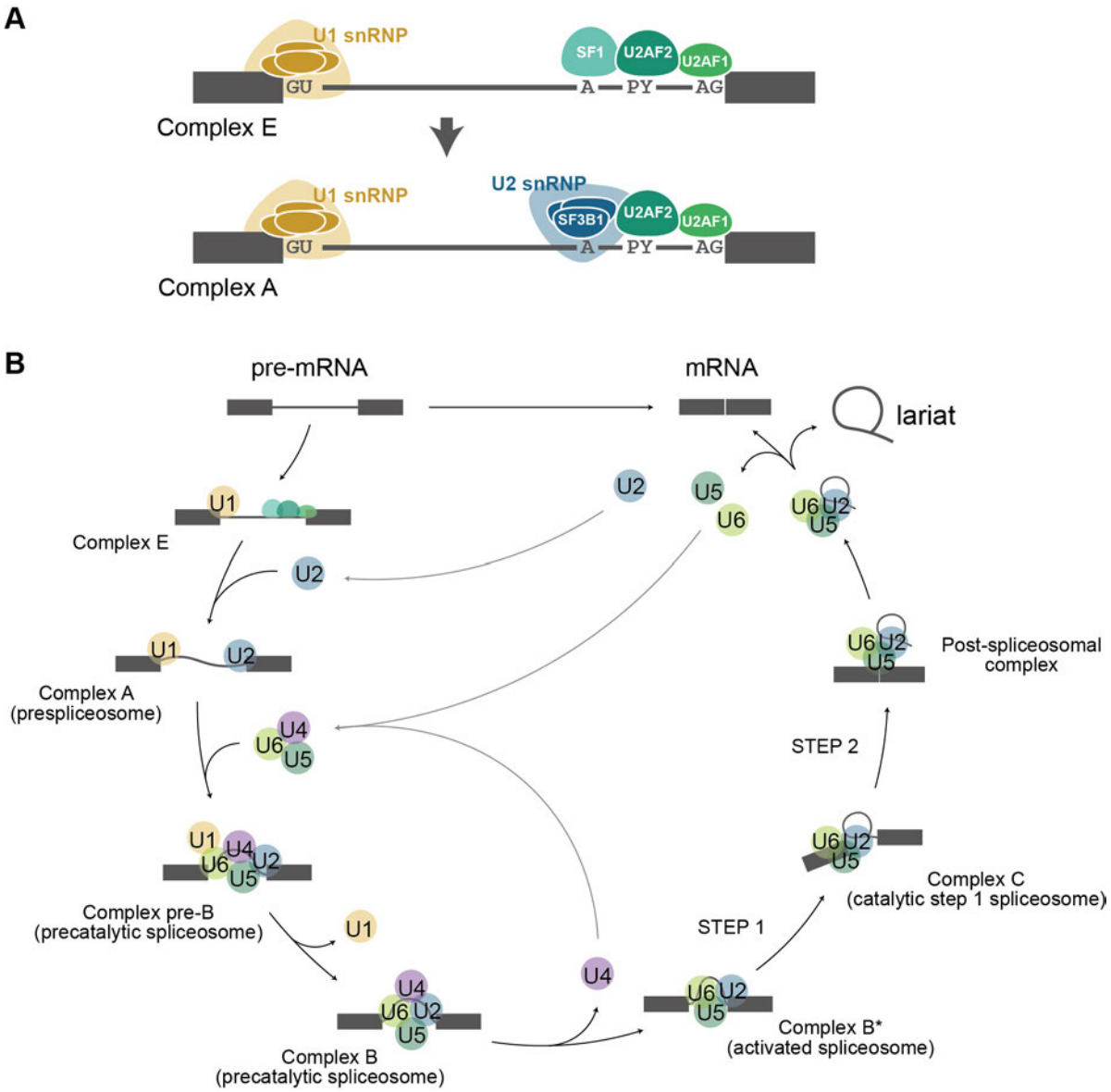


Figure 2: Molecular mechanisms of splicing. (A) Splice site definition is facilitated by U1 snRNP binding to the 5'ss and U2AF1 and U2AF2 binding to the 3'ss and the polypyrimidine tract (PY), respectively. They recruit the branch point binding protein SF1 to form complex E. Now U2 snRNP is recruited to the branch point adenosine, where SF3B1 substitutes SF1, forming complex A. (B) Full splicing circle involves formation of complex E and complex A. Subsequently, the tri-snRNP is recruited and activated by releasing U4 snRNP. U5 snRNA catalyzes the branch point attack and lariat formation. After exon ligation, the lariat is released and the spliceosome disassembles from the exon-exon junction

Subsequently, the complex A is formed by recruiting U2 snRNP to the 3'ss, with U2 component SF3B1 replacing SF1 at the branch point (**Figure 2**A and B). Next, the tri-snRNP enters the spliceosome, with U4 snRNA and U6 snRNA engaging in extensive base pairing. The 5'ss is transferred from the U1 snRNA to the U6 snRNA (complex pre-B). While U1 leaves the complex, U6 snRNA unwinds from the U4 snRNA, enabling U6 snRNA to also associate with part of the U2 snRNA (complex B). U6 snRNA is now attached to both the 5'ss and the branch point (BP) and forms the catalytic center of the spliceosome, rendering it a ribozyme (complex B*). The BP adenosine branches now to the 5'ss, forming the lariat, which can now leave the active site (complex

C). The cleaved 5' exon remains in the active site, attacking the 3'ss to form an exon-exon ligation (post-spliceosomal complex). The lariat and the splicing proteins dissociate from the mRNA and thus ending the intricate choreography of the spliceosome (Kastner et al. 2019; Wilkinson, Charenton, and Nagai 2020) (**Figure 2**B).

1.2.2. Splicing regulation and specificity

Due to the complexity of splicing and the high number of involved proteins, splicing requires meticulous regulation. Surprisingly, pre-mRNA itself offers little information to the spliceosome. Compared to lower eukaryotes, humans have degenerate and fewer splice sites (Corioni et al. 2011). For instance, while the yeast branch point has a conserved sequence (UACUAAC), the human branch point is set in a less defined sequence (YNYURAY) (Gao et al. 2008). Despite this apparent loss of information, this situation actually harbors immense regulatory possibilities. The sparse information calls for a more fine-tuned, sophisticated network of decision-making, influencing splicing speed, location, and ultimately, dictating splicing outcome.

The sequence variety of *cis*-regulatory elements allows them to differ in strength or length. For example, a py-tract can be shorter or longer, enabling U2AF2 to bind slowly or rapidly, respectively. Splicing is very dynamic and happens in only a few minutes, so time is of the essence and a missed 3'ss recognition can lead to exon skipping. Additionally, regulatory sequences on the pre-mRNA, like intronic splicing silencers (ISS) and exonic splicing enhancers (ESE) evolved in higher eukaryotes to influence splicing by recruiting hnRNPs or serine-arginine (SR)-rich proteins. SR-proteins are generally known to promote exon inclusion while hnRNPs favor exon skipping. However, the presence of both components is vital for the splicing decision, thus adding an intricate balance to the spliceosomal machinery (Nilsen and Graveley 2010).

Splicing is rarely performed on a fully transcribed pre-mRNA, but often occurs co-transcriptionally, meaning that intron excision can occur before the pre-mRNA – or even the downstream exon – is completely transcribed. This mechanistically links splicing to transcription, introducing another layer of complexity in splicing decision-making and regulation. Introns act independently from each other and one transcript can have both co- and post-transcriptionally excised introns (Bentley 2014; Tellier, Maudlin, and Murphy 2020). How much of splicing is performed before or after transcription termination, has not been answered yet. But the effect of co-transcriptional splicing is quite evident, as transcription rate can determine the folding of the pre-mRNA and therefore the accessibility of the splice sites to the spliceosome (Nilsen and Graveley 2010). In addition, recent studies have found that U1 snRNP is tethered to the C-terminal domain (CTD) of RNA Polymerase II (PolII), enabling U1 snRNP to influence transcription speed and to directly bind to the 5'ss at the moment of its emergence (Mimoso and Adelman 2023; S. Zhang et al. 2021). Additionally, the U2 snRNP

positively influences transcriptional speed, as paused PolII relies on a functional spliceosome to be released and continue with transcription elongation (Caizzi et al. 2021). Transcription speed and therefore splicing outcome can also be determined by epigenetic marks. It has been shown that the exonic sequences in the genome are more likely to be histone-bound. If the histones harbor H3K36me3, PolII slows down, allowing splice site recognition. Alternative exons often lack H3K36me3, therefore they are being rapidly passed by the polymerase and more likely skipped (Nilsen and Graveley 2010).

Interestingly, intron length also profoundly affects splicing decisions. Introns can differ in length, ranging from only 30 nucleotides (nt) up to $1x10^6$ nt, however most frequently, human introns are approximately 1500 nt long (Dvorak, Hanicinec, and Soucek 2023; Piovesan et al. 2015, 2019). Longer introns express a differential GC content, which means that the GC content in introns is lower than in the exons. Exons flanked by short introns do not only have a higher GC content, but also the exon-intron GC ratio is leveled (Amit et al. 2012). Compared to short introns, long introns are rather spliced post-transcriptionally, after being transported from the nuclear center to the periphery. Intriguingly, intron length also determines splicing outcome. For long introns, exon skipping is more common, while short introns are more prevalently retained (Fox-Walsh et al. 2005; Tammer et al. 2022). This pattern holds true both for individual introns, but also on a species level (Keren, Lev-Maor, and Ast 2010).

1.3. Splice site recognition across evolution

Especially for long introns, it is worthwhile to have a look at splicing throughout the phylogenetic tree. Though splicing can be traced back to the last common eukaryotic ancestor, splicing differs immensely from yeast to human (Ule and Blencowe 2019). For an intron to be excised, the boundaries of exons and introns must be defined. The higher GC content of the exon compared to its flanking introns is believed to be the main recognition element to define introns and exons (Keren, Lev-Maor, and Ast 2010). The splicing machinery assembles at the exon-intron boundaries and bridge the splice sites. This happens by either intron definition or exon definition. During intron definition, the splice sites are bridged over the intron (**Figure 3**A). Exon definition is a two-step process, in which the definition is initially performed over the exon and subsequently over the intron (**Figure 3**B). Generally, intron definition is considered to be the ancient recognition mechanism, which is still employed by lower eukaryotes, while exon definition evolved later and is believed to be the main course of action in higher eukaryotes (Keren, Lev-Maor, and Ast 2010). The two modes of splice site bridging puts evolutionary pressure on the length of the recognized elements. To be able to bridge over an intron, introns must remain short. On the other hand, when splice sites are bridged over the exon, the exon must remain short, however, intron length is not put under the same

constraints. Therefore, in animals, we prevalently find long introns, interspersing rather short exons (De Conti, Baralle, and Buratti 2013).

Though the initial definition of splice sites is performed by exon definition in mammals, eventually, two consecutive exons have to be brought into close proximity to be ligated together. To do so, different species have different approaches. For example, *D. melanogaster* solve this problem by recursive splicing (Pai et al. 2018). Here, non-constitutive splice sites are used to excise small parts of the intron, shorten it gradually, until all of the intron is excised. In humans, existence of recursive splicing is still debated (X.-O. Zhang et al. 2018). Hoppe and colleagues provided us a deeper understanding of recursive splicing only this year and show that indeed, recursive splicing is more prevalent in humans than previously expected and that these recursive splice sites have the power to influence exon inclusion (Hoppe, Udy, and Bradley 2023).

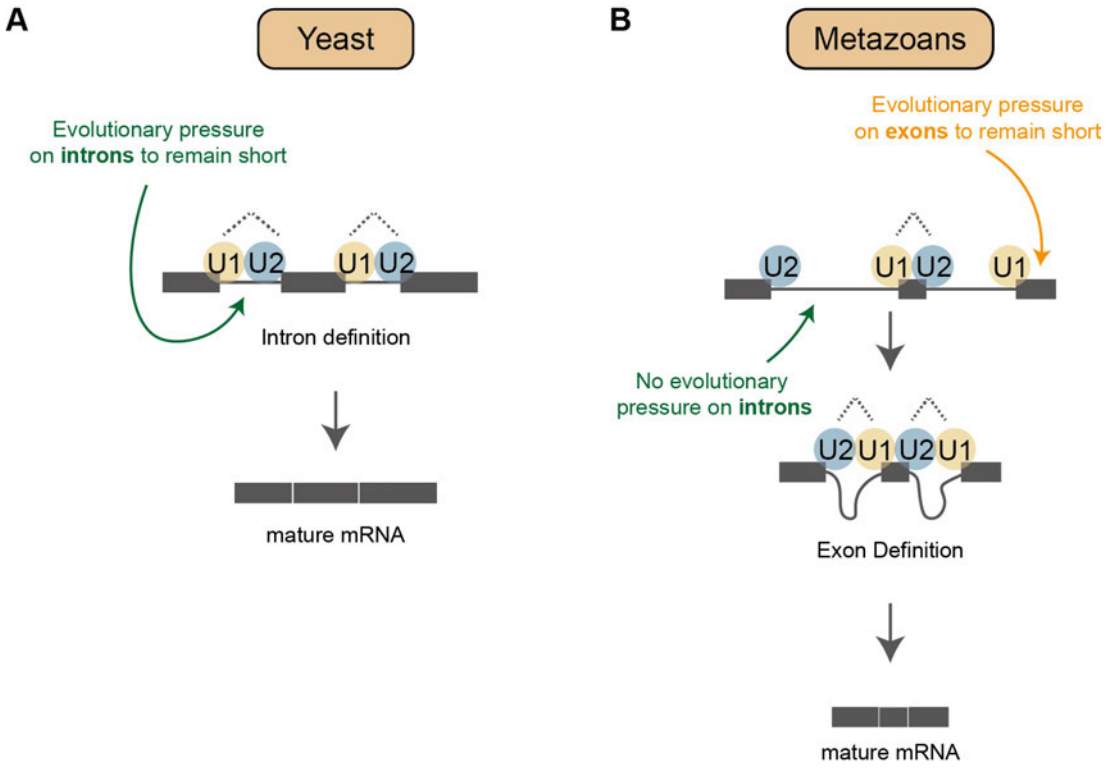


Figure 3: Splicing in evolution. (A) In yeast, splice site definition is facilitated over the intron (intron definition). Therefore, there is evolutionary pressure on the introns to remain short. (B) In contrast to yeast, metazoans facilitate exon definition, where the splice sites are first connected over the intron, then over the intron. Here, the evolutionary pressure is on the introns to exons to remain short. There is no evolutionary pressure on the introns.

1.4. Splicing in health, disease and cancer treatment

1.4.1. Splicing-related impairments

The importance of splicing becomes particularly apparent when splicing is impaired, for example in cancer. Defective splicing is a hallmark of cancer, but is often underappreciated in treatment development. In a comprehensive study over 32 cancers with over 8000 tumor samples, it has been shown that malignant tissue expresses thousands of cancer-specific splicing variants that are absent in non-malignant tissue (Kahles et al. 2018). In a few cases, these mRNA variants can even lead to treatment resistance but more importantly, these unique markers could also serve as neo-antigens for targeted cancer therapy (Frankiw, Baltimore, and Li 2019).

Often, cancer-specific splicing changes occur due to misregulated or mutated splicing factors. A study of 33 cancer types and more than 10 000 samples revealed over 119 splicing factor genes, harboring recurrent mutations in both solid and hematological cancers (Seiler et al. 2018). Amongst them, the gene of the U2 snRNP protein SF3B1 is one of the most commonly mutated. In a subtype of myelodysplastic syndrome (MDS), 65-83% of patients show a mutated *SF3B1* gene (Papaemmanuil et al. 2011; Yoshida et al. 2011). Mutations in *SF3B1* can cause cryptic splice site usage and therefore aberrant splicing isoforms (Alsafadi et al. 2016). Apart from tumors, there are also splicing impairments that are associated with neurological, immunological and even infectious diseases (Y. Zhang et al. 2021). For example, in neurodegenerative disorder Huntington's disease (HD), SRSF6 is upregulated, highly phosphorylated and accumulates in inclusion bodies. This leads to an aberrant splicing pattern that might drive HD pathology (Fernández-Nogales et al. 2016).

One of the best-studied splicing-related diseases is spinal muscular atrophy (SMA). In 90% of all cases, SMA arises from either deletion or mutation of the *SMN1* gene. SMN2, though being a close homolog of SMN1, cannot compensate for the loss of SMN1 due to skipping of the alternative exon 7. Surprisingly, recent advances show that antisense oligo (ASO)-mediated inclusion of *SMN2* exon 7 makes SMN2 a viable substitute for SMN1 and thereby alleviates SMA symptoms (R. N. Singh, Seo, and Singh 2020; R. N. Singh and Singh 2018).

1.4.2. Splicing affects CART-19 therapy in B-ALL patients

Splicing does not only affect the health of an organism, but can also influence treatment of diseases like cancer. This is also the case for B cell acute lymphoblastic leukemia (B-ALL) patients. B-ALL can occur in both children and adults. However, it is more common in young individuals, as 60% of all B-ALL patients are under the age of 20 (Jabbour et al. 2015). B-ALL arises due to genetic alterations in lymphoid precursor cells, the lymphoblasts (Bertrand et al. 2001), leading to an accumulation of immature, non-functional lymphoblasts (Terwilliger and Abdul-Hay 2017). In the blood, the increased number of blasts leads to the reduced levels of other blood cells like erythrocytes,

platelets and other leukocytes, ultimately causing anemia. Therefore, patients suffering from B-ALL often experience symptoms like shortness of breath, bruising and frequent infections. As the most common cancer type in children, treatment of B-ALL is of major interest. Chemotherapy leads to good results in children, but the 5-year survival rate for adults is only 30-40% (Jabbour et al. 2015; Narayanan and Shami 2012). Therefore, B-ALL patients are in need of alternative treatment methods.

One of the most recent advances in cancer treatment is a cell-based approach, called CART cell therapy. Here, T cells are obtained from the patient and trained to recognize antigens presented on cancer cells. This enables the patient's own immune system to combat the tumor. In detail, after T cell isolation, the T cells are cultured *ex vivo* and via viral transfection, are equipped to produce chimeric antigen receptors (CAR). Those consist of the single-chain variable fragment (scFv) of an antibody detecting the epitope of interest, and the intracellular signaling domain of the T cell receptor (Sterner and Sterner 2021). In the case of B-ALL, the CAR recognizes the B cell-specific surface marker CD19, which is also expressed on the faulty leukemia lymphocytes. The engineered CART-19 cells are expanded in cell culture and transplanted back to the patient. Here, the T cells recognize the malignant B cells by the epitope of interest, CD19, thereby activating the release of cytolytic molecules, and killing the detected cancer cell. In addition, the release of cytokines alert other effectors of the immune system, aiding in the eradication of the cancer cell that is now recognized as a foreign invader (Maude et al. 2014). The CART-19 cell is the first CART cell therapy that was released to the market (Sterner and Sterner 2021).

However, still, up to half of patients treated with CART-19 relapse (Wudhikarn et al. 2021). In 40-60% of the patients, it was found that the malignant B cells lost the CART-19 epitope (Maude et al. 2016; Orlando et al. 2018). Interestingly, this loss of epitope has been recurrently linked to alternative splicing. By skipping *CD19* exon 2, the resulting truncated protein is unable to locate to the surface and instead, remains cytoplasmic. Thereby, the CART-19 cells is unable to recognize their target (Bagashev et al. 2018; Sotillo et al. 2015). This is the most common mode of CART-19 evasion seen in B-ALL patients, but there are also other mechanisms e.g. *CD19* intron 2 retention, or skipping of both exon 5 and 6, both leading to a premature termination codon (PTC) (Asnani et al. 2020). How malignant lymphoblasts induce alternative splicing, and thereby epitope loss and CART-19 evasion, remains to be shown.

1.5. FUBP1 – a multifaceted protein

1.5.1. FUBP1 regulates U2AF2 binding to the py-tract and is of particular interest

With the advancements in high-throughput pipelines and cryo-EM, detailed knowledge about splicing mechanisms could have been uncovered. Numerous outstanding reviews (Wahl and

Lührmann 2015; Wahl, Will, and Lührmann 2009; Wilkinson, Charenton, and Nagai 2020) carefully describe the ever-increasing amount of information. However, mysteries persist, particularly in target recognition and spliceosome assembly during RNA's dynamic life cycle.

Recently, [redacted] lab used machine learning paired with high-throughput experiments to reveal regulators of U2AF2, a key component in 3' splice site recognition. Among them, Far-Upstream-Element binding protein 1 (FUBP1, previously FBP1) stood out, enhancing U2AF2 binding to the py-tract (Reymond Sutandy et al. 2018). This finding is particularly intriguing due to the role of FUBP1 as a transcription factor for the proto-oncogene MYC and its strong link to various cancer types. While the role of FUBP1 as a DNA-binding protein is well established, its role in RNA, particularly splicing, remains elusive.

1.5.2. FUBP1 is a DNA- and RNA-binding protein

FUBP1 belongs to the FUBP protein family, consisting of three members. FUBPs play crucial roles throughout the RNA life cycle, participating in processes like mRNA synthesis or degradation, RNA transport, or translational regulation (Briata et al. 2016; Chung et al. 2006; J. Zhang and Chen 2013). All three proteins share high sequence homology and contain four K-homology domains that can bind nucleic acid sequences (Debaize and Troadec 2019). Both FUBP1 and FUBP3 exhibit the ability to bind both DNA and RNA, KHSRP (also known as FUBP2) is known to solely bind RNA. KHSRP regulates post-transcriptional processes in the nucleus, while FUBP3 is involved in translation regulation in the cytoplasm (Briata et al. 2016; Gau et al. 2011). Notably, FUBP1 has been extensively studied as a transcription factor and regulator of MYC (Duncan et al. 1994).

FUBP1 is a 68-kDa protein with an N-terminal helix, three nuclear localization signals (NLS) and four KH domains, and two C-terminal domains of yet unknown function (**Figure 4**A). The three NLS of FUBP1 ensure its proper localization in the nucleus. However, when fused to GFP, the N-terminal NLS of FUBP1 is sufficient to drive nuclear localization of GFP (He, Weber, and Levens 2000). Under stress like heat shock, viral infection and oxidative stress, FUBP1 is cleaved within the N-terminal NLS by caspase-3 and -7, and subsequently exported to the cytoplasm (Jang et al. 2009; Thiede et al. 2001). The gene transcribing FUBP1, located on the reverse strand of chromosome 1 (1p31.1), displays a high degree of conservation throughout the tree of life, with 90% sequence homology throughout mammals and notable 73% with *Danio rerio* (Chung et al. 2006; Davis-Smyth et al. 1996; Debaize and Troadec 2019). FUBP1 is ubiquitously expressed, with a particular enrichment in unproliferated myeloid cells (Avigan, Strober, and Levens 1990; Davis-Smyth et al. 1996). It comes in two isoforms, with isoform 1 containing an additional in-frame exon within the 5'-coding region. Functionally, FUBP1 is a multifaceted protein with the ability to bind both DNA- and RNA, making it a "master regulator" of gene expression across various intricate networks

(Debaize and Troadec 2019). While its role in DNA is well established, its function as an RBP has only come under scrutiny in the last decade.

1.5.3. FUBP1 regulates cell survival by its DNA-binding functions

FUBP1 was first discovered in its function of binding to the Far-Upstream binding site (FUSE), a distal sequence within the *MYC* promoter, found to regulate the expression of the general transcription factor and proto-oncogene MYC (Duncan et al. 1994). Because MYC is involved in a multitude of processes, its proper expression is essential for the cell, and its misregulation often leads to cancer (Duffy et al. 2021). Consequently, misregulation or mutations in FUBP1 also lead to a variety of cancers, of which low grade lymphoma (LGG) is the most common (Debaize and Troadec 2019; Seiler et al. 2018). Therefore, fine-tuning of *MYC* regulation by FUBP1 is a vital mechanism to maintain cell health.

Mechanistically, FUBP1 activates *MYC* expression by binding to FUSE within the *MYC* promoter and thereby interacting with the basal transcription factor TFIIH, enhancing the helicase activity of TFIIH. On the other hand, the FUBP1 interacting repressor (FIR) competes for FUSE binding, creating a FUSE/FUBP1-FIR complex. Herein, FIR directly interacts with FUBP1, and then replaces it. This inhibits the helicase function of TFIIH, leading to downregulation of *MYC* expression (**Figure 4**B, left). Consequently, misregulation of FUBP1 disrupts both up- and downregulation of *MYC* expression, illustrating the sensitivity of the fine-tuned regulation of *MYC* (Debaize and Troadec 2019).

Within FUBP1, KH3 and KH4 are sufficient to bind FUSE *in vitro* (Duncan et al. 1994). KH4 and KH3 directly interact with the short sequences 5′-TATTCC -3′ and 5′-ATTTTT -3′, respectively, with a six-nucleotide separation believed to be unbound by FUBP1 (DT et al. 2002). Beyond FUSE, FUBP1 has shown binding to other so-called “FUSE-like” sequences, influencing the expression of various genes such as *USP29*, c-*KIT*, *CDKN1A*, and *BIK* both positively and negatively (Debaize and Troadec 2019). While these bound motifs are characterized by AT- or GT-rich stretches, they also display variations, indicating the flexible binding capacity of FUBP1. How FUBP1 gains specificity, however, has not been understood so far.

Moreover, FUBP1 plays a multifaceted role in cell processes beyond *MYC* regulation. It regulates cell cycle progression by enhancing Cyclin D1 and Cyclin D2 expression while inhibiting the expression of the cell cycle arrest inducer p21, further favoring cell cycle progression (Atanassov and Dent 2011; Rabenhorst et al. 2009, 2015). By regulating Cyclin D1 and Cyclin D2 expression, FUBP1 regulates cell cycle progression. Inversely, FUBP1 inhibits expression of cell cycle arrest inducer p21, further advantaging cell cycle progression (Gartel, Serfas, and Tyner 1996). Additionally, FUBP1 promotes cell survival by repressing transcription of pro-apoptotic proteins like

NOXA or TNFA in Hep3B cells (Malz et al. 2014). In summary, FUBP1 serves not only as a transcription factor but also as a vital tumor regulator, assuring cell proliferation and survival.

1.5.4. FUBP1 as an RNA-binding protein

In the last decade, extensive research has begun to unveil the multifaceted role of FUBP1, extending beyond its initial identification as a transcription factor. It has been shown that its KH domains enable it to bind RNA, although the precise mechanisms underlying the influence of FUBP1 on the RNA life cycle remains elusive. Recent studies have linked FUBP1 to various post-transcriptional regulatory processes, including translational repression, mRNA degradation, splicing, and viral replication (Debaize and Troadec 2019).

FUBP1 exhibits a preference for GU-/AU-rich sequences in RNA, similar to its binding affinity for GT/AT-rich sequences in DNA (Debaize and Troadec 2019; Ni, Knapp, and Chaikuad 2020). Moreover, it has been shown that FUBP1 also targets AU-rich elements (AREs) present in the 3'UTRs of short-lived RNA species (Sully et al. 2004; W. Zheng et al. 2016). By binding to the 3'UTR, and in some cases specifically AREs in 3'UTRs, FUBP1 negatively regulates protein expression, either by promoting mRNA degradation or inhibiting translation of its target transcripts (Irwin et al. 1997; W. Zheng et al. 2016). Conversely, in some instances, FUBP1 enhances protein translation by binding to the Internal Ribosome Entry Site (IRES) in the 5'UTR of specific mRNAs, such as *p27* mRNA in MCF7 cells (Yuhuan Zheng and Miskimins 2011). These findings highlight that FUBP1 is a versatile regulator, capable of both upregulating and downregulating post-transcriptional processes (**Figure 4**B, right).

One of the most intriguing roles of FUBP1 is its involvement in splicing regulation (**Figure 4**B, bottom). As early as 2002, Rappsilber et al. demonstrated the association of FUBP1 with the spliceosome through a massive pulldown experiment (Rappsilber et al. 2002). Notably, FUBP1 appears to have a context-dependent effect on splicing, exhibiting both enhancing and inhibitory influences. Binding to AU-rich sequences within an exonic splicing silencer located in exon 10 of the *triadin* mRNA leads to the suppression of the second esterification reaction, thereby inducing skipping of exon 10. Similarly, FUBP1 promotes exon skipping in *ACLY* exon 14 and caspase 9 exons 4-7. In contrast, FUBP1 promotes exon inclusion of PTBP2 exon 10 and *ENAH/MENA* exon 11 (Li 2013). Additionally, in cooperation with its homolog KHSRP, FUBP1 contributes to exon skipping in the *SMN2* gene, resulting in the shortening of the *SMN* protein, rendering it non-functional (J. Wang, Schultz, and Johnson 2018). Remarkably, FUBP1 demonstrates a pivotal role in the regulation of *MDM2* exon 11, favoring constitutive splicing by skipping exons 4-11, thereby suppressing the alternative isoform MDM2-B, which lacks the interaction interface to p53 (Jacob et al. 2014). The accumulation of MDM2-B has been observed in several cancer types, underscoring

FUBP1's tumor-suppressor function. FUBP1 also plays a critical role in *DMD* exon 39, where it is binding to an intronic splicing enhancer in intron 38, approximately 80 nt upstream of *DMD* exon 39, ensures normal exon inclusion, essential for the proper expression of dystrophin protein. Dystrophin misregulation has been associated with muscular dystrophies and cardiomyopathy (Miro et al. 2015). Furthermore, FUBP1 facilitates the inclusion of *LSD1* exon 8a, promoting terminal neuronal differentiation and functioning as a tumor suppressor (Hwang et al. 2018).

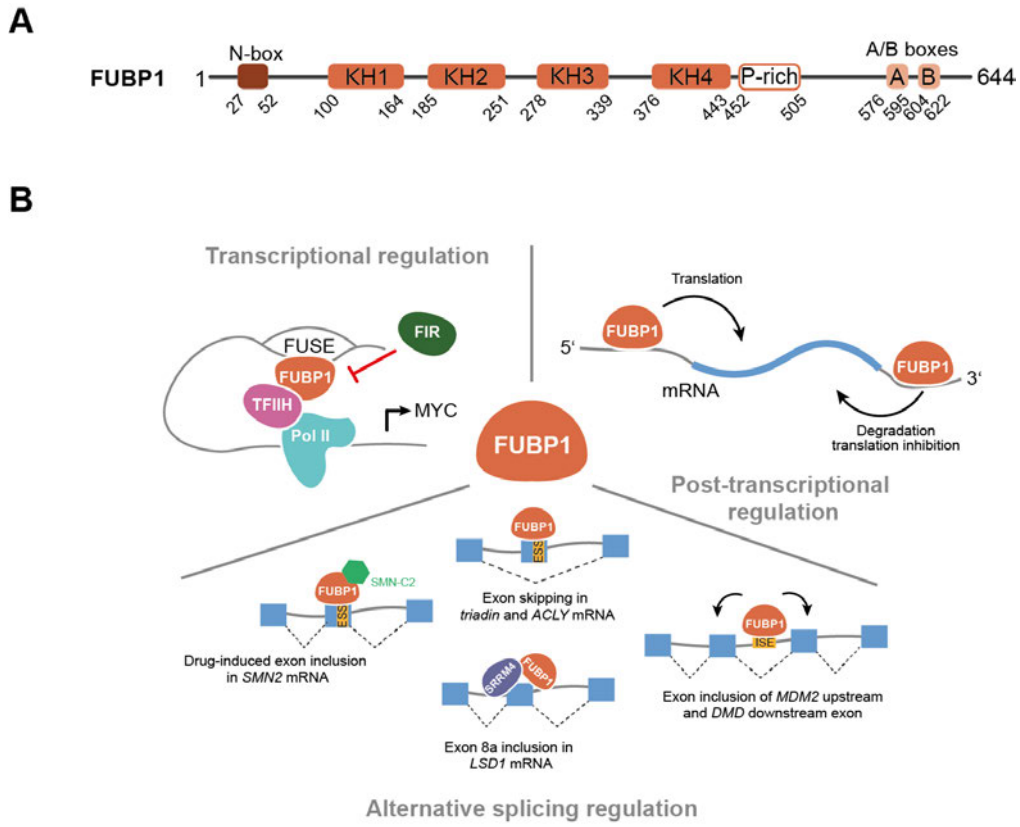


Figure 4: Characterization of FUBP1. (A) Domain structure of FUBP1. FUBP1 consists of an N-terminal helix (N-box), four KH domains that can bind single stranded nucleic acids, a proline-rich sequence, and two FUBP-family specific motifs (A/B boxes). (B) Gene regulatory roles of FUBP1. FUBP1 acts as a transcription factor of MYC by binding to the FUSE sequence. FUBP1 can influence post-transcriptional regulation by binding to the 5' and 3' UTR. FUBP1 binds both to intronic and exonic sequences, promoting both exon skipping and inclusion.

While these regulatory roles have been well documented for specific loci, the global impact of FUBP1 in splicing regulation remains an ongoing area of exploration. Additionally, in flies, FUBP1 is known to interact with the U1 component SNRNPK (also known as U170K), suggesting its involvement in intricate splicing networks (Ignjatovic et al. 2005; Labourier, Adams, and Rio 2001). Together, though our present knowledge about FUBP1 as an RNA-binding protein indicate an important role in splicing, further research is needed to understand the detailed mechanisms in which FUBP1 is involved.

2. Aim of the work

FUBP1 is mainly known to be a transcription factor. Its role as a regulator of MYC expression and its interaction to FIR are well understood. However, recent reports uncovered that FUBP1 influences exon inclusion and skipping events. In a previous study from [redacted] it was observed that FUBP1 influences U2AF2 binding to the py-tract. However, the global impact of FUBP1 on splicing is not understood. Therefore, we aim to unravel the functional relevance of FUBP1 in splicing.

To start, we investigate whether FUBP1 binds to pre-mRNA via iCLIP, in which we crosslink FUBP1 to the RNA. Subsequent NGS will reveal whether and where FUBP1 binds the RNA, as well as give insight into an FUPB1-specific motif. Because iCLIP can be biased in nucleotide composition, we will collaborate with the group of [redacted] to complement our results with an NMR-based approach called scaffold-independent analysis. Next, we want to investigate whether FUBP1 interacts with other splicing components, especially U2AF2. To do so, together with the group of [redacted], we will employ a BRET-based reporter assay, called LuTHy. Again, our obtained knowledge about the protein interactions will be deepened by NMR-based interaction assays, which will be executed by the [redacted].

To understand which pre-mRNAs are targeted by FUBP1, we will create two *FUBP1* mutant cell lines by CRISPR/Cas9 editing. One cell line will still express *FUBP1*, but will be functionally impaired. In the other cell line, *FUBP1* will be completely knocked out. We will perform RNA-seq to investigate which splicing events suffer most from the *FUBP1* mutations. Further, we will create a minigene of an FUBP1 target, in which we can selectively delete FUBP1-specific features and motifs. This will help us understand, how FUBP1 mechanistically influences splicing decisions. To understand whether FUBP1 also influences splicing kinetics, we will perform nascent RNA-seq, which will show us which type of RNA will be spliced slower upon *FUBP1* depletion.

Lastly, we aim to understand the role of FUBP1 throughout evolution. To this end, we collaborate with [redacted] to search for similarities and differences in *FUBP1* genes across the phylogenic tree. With this, we try to unravel how FUBP1 interacts with its targets and with the spliceosome, and link those findings its role in splicing regulation.

In addition to understand the role of FUBP1 in splicing, we also investigate regulation of *CD19* mRNA splicing. B-ALL patients treated with CART-19 and relapsed, often exhibit *CD19* mis-splicing. We want to understand, which *cis*-regulatory elements influence *CD19* mRNA splicing. Therefore, a high-throughput mutagenesis assay will be performed, which will be verified by semi-quantitative RT-PCR of the observed mutations. In addition, short hairpin RNA (shRNA)-induced KD of splicing factors will unravel the influence of *trans*-acting factors on *CD19* mRNA splicing.

3. High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling *CD19* splicing and CART-19 therapy resistance

3.1. Abstract

B cell acute lymphoblastic leukemia (B-ALL) is a cancer type with high prevalence in children, but can be treated with a CART-19 therapy. However, a high percentage of patients relapse due to loss of epitope. In this study, we set out to understand the potential mechanisms of epitope loss of CD19. We used a high-throughput mutagenesis approach to identify candidate mutations that alter *CD19* splicing. We found approx. 200 mutations resulting in over 100 cryptic splice isoforms. Cloning of selected mutations into a *CD19* minigene and subsequent RT-PCR validated these results. In addition, by shRNA-induced KD of splicing proteins, we could show that the loss of *trans*-acting factors like SF3B4 and PTBP1 favor *CD19* mis-splicing. With this study, we offer a comprehensive overview of potential sources of CD19 epitope loss, which can lead to B-ALL relapse after CART-19 therapy.

3.2. Zusammenfassung

Akute lymphoblastische B-Zellleukämie (B-ALL) ist eine Krebsform mit hoher Prävalenz unter Kindern, kann aber mit einer CART-19 Therapie behandelt werden. Dennoch erleidet eine hohe Anzahl an Patienten einen Rückfall durch den Verlust des Epitops. In dieser Studie wollen wir die möglichen Mechanismen des CD19-Epitopverlust untersuchen. Wir verwendeten eine Hochdurchsatzmutagenese, um potentielle Mutationen zu identifizieren, die das Spleißen von *CD19* verändern. Wir konnten ca. 200 Mutationen finden, die in über 100 kryptische Isoformen resultiert. Klonierungen ausgewählter Mutationen in ein *CD19* Minigen und anschließende RT-PCR konnte diese Ergebnisse reproduzieren. Des Weiteren konnten wir durch shRNA-induzierten KD von Spleißproteinen zeigen, dass der Verlust von *trans*-agierende Faktoren wie SF3B4 und PTBP1 das *CD19* Misspleißen favorisieren. Mit dieser Studie bieten wir einen umfassenden Überblick über die möglichen Quellen des CD19-Epitopverlusts, die zu einem Rückfall der B-Zellleukämie nach CART-19-Therapie führen kann.

3.3. Statement of contribution

The high-throughput mutagenesis of the *CD19* minigene yielded over 200 different mutations that lead to cryptic isoforms. To validate these results, together with [redacted], we chose different mutations to verify. For half of the mutations, I designed and executed site-directed mutagenesis to create minigenes with the determined mutations. I performed cloning and validation of the sequences

with Sanger sequencing, while [redacted] performed transfection and RT-PCR, as well as the evaluation.

In addition, together with [redacted], we performed parts of the revision. The reviewer asked whether the shRNA-induced *PTBP1* KD in HEK293 cells changed the CD19 cell surface expression in a similar fashion as has been described for the siRNA-induced *PTBP1* KD in P493-6 and MHHCALL4 cells. With RT-PCR, we wanted to understand which cryptic isoforms are generated in the shRNA-induced KD. While [redacted] performed transfection and induction, I performed the RT-PCR and evaluation. In addition, [redacted] and I performed flow cytometry to observe surface CD19 on the transfected and induced cells. However, the results were not included. Finally, I reviewed the manuscript.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-31818-y OPEN

High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling *CD19* splicing and CART-19 therapy resistance

Mariela Cortés-López[1,14], Laura Schulz[1,14], Mihaela Enculescu[1,14], Claudia Paret[2,3,4], Bea Spiekermann[1], Mathieu Quesnel-Vallières[5,6], Manuel Torres-Diz[7], Sebastian Unic[8], Anke Busch[1], Anna Orekhova[1], Monika Kuban[8], Mikhail Mesitov[1], Miriam M. Mulorz[1], Rawan Shraim[7,9], Fridolin Kielisch[1], Jörg Faber[2,3,4], Yoseph Barash[5], Andrei Thomas-Tikhonenko[7,10], Kathi Zarnack[11,12]✉, Stefan Legewie[1,8,13]✉ & Julian König[1]✉

Following CART-19 immunotherapy for B-cell acute lymphoblastic leukaemia (B-ALL), many patients relapse due to loss of the cognate CD19 epitope. Since epitope loss can be caused by aberrant *CD19* exon 2 processing, we herein investigate the regulatory code that controls *CD19* splicing. We combine high-throughput mutagenesis with mathematical modelling to quantitatively disentangle the effects of all mutations in the region comprising *CD19* exons 1-3. Thereupon, we identify ~200 single point mutations that alter *CD19* splicing and thus could predispose B-ALL patients to developing CART-19 resistance. Furthermore, we report almost 100 previously unknown splice isoforms that emerge from cryptic splice sites and likely encode non-functional CD19 proteins. We further identify *cis*-regulatory elements and *trans*-acting RNA-binding proteins that control *CD19* splicing (e.g., PTBP1 and SF3B4) and validate that loss of these factors leads to pervasive *CD19* mis-splicing. Our dataset represents a comprehensive resource for identifying predictive biomarkers for CART-19 therapy.

[1] Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany. [2] Department of Pediatric Hematology/Oncology, Center for Pediatric and Adolescent Medicine, University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany. [3] University Cancer Center (UCT), University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany. [4] German Cancer Consortium (DKTK), site Frankfurt/Mainz, Germany, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [5] Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA. [6] Department of Biochemistry and Biophysics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA. [7] Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [8] Department of Systems Biology, Institute for Biomedical Genetics (IBMG), University of Stuttgart, Allmandring 30E, 70569 Stuttgart, Germany. [9] Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA. [10] Department of Pathology & Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA. [11] Buchmann Institute for Molecular Life Sciences (BMLS), Max-von-Laue-Str. 15, 60438 Frankfurt, Germany. [12] Faculty Biological Sciences, Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt, Germany. [13] Stuttgart Research Center for Systems Biology (SRCSB), University of Stuttgart, Stuttgart, Germany. [14] These authors contributed equally: Mariela Cortés-López, Laura Schulz, Mihaela Enculescu. ✉email: kathi.zarnack@bmls.de; legewie@ibmg.uni-stuttgart.de; j.koenig@imb-mainz.de

B-cell acute lymphoblastic leukemia (B-ALL) is a haematologic malignancy that causes a significant number of childhood and adult cancer deaths. During CART-19 immunotherapy, chimeric antigen receptor-armed autologous T-cells (CARTs) are engineered to target the surface antigen CD19 on B cells by linking the single-chain variable fragment (scFv) of an anti-CD19 antibody to the intracellular signalling domain of the T-cell receptor[1]. Upon CD19 recognition, the chimeric antigen receptors activate the cytotoxic T-cells to attack the tumour cells. CART-19 therapy was recently approved for the treatment of paediatric B-ALL in the US and Europe, achieving initial remission rates up to 90%[2]. This indicates that B-ALL cells at initial screening are typically CD19-positive. Unfortunately, up to 50% of children relapse under CART-19 therapy, and response rates are even worse in adults[3,4]. Several studies reported that in 40–60% of relapse cases, the cancerous B cells become invisible to the CARTs because they lose expression of the CD19 epitope (CD19-negative)[5–8]. This recurrently involves alternative splicing of the *CD19* pre-mRNA[9–11].

Splicing involves the excision of introns and the joining of exons by the spliceosome to generate mature mRNAs. In alternative splicing, certain exons can be either included or excluded ("skipped"), resulting in different transcript isoforms. The splicing outcome at each exon is controlled by a large set of *cis*-regulatory elements in the RNA sequence which are recognised by *trans*-acting RNA-binding proteins (RBPs) that guide the spliceosome activity. It is increasingly recognised that widespread alterations in splicing are a molecular hallmark of cancer and often contribute to therapeutic resistance (reviewed in ref. [12]). For instance, intron retention, i.e., the failure to remove certain introns, often disrupts the open reading frame (ORF) with premature termination codons (PTCs) and thereby compromises the expression of the encoded proteins. Consistent with the widespread splicing changes, cancer-causing driver mutations frequently occur in splice-regulatory *cis*-elements, and many splicing factors have oncogenic properties, being commonly mutated or dysregulated in cancer[12–14].

Multiple alternative splicing events in *CD19* mRNA have been described to interfere with CART-19 therapy[9,11,15–18]. Most prominently, skipping of exon 2 results in a truncated CD19 protein which is no longer presented on the cell surface and hence fails to trigger CART-19-mediated killing[9,15]. In addition, it was reported that relapsed patients showed retention of intron 2 which introduces a PTC, thereby disrupting CD19 expression[11]. Similarly, simultaneous skipping of exons 5 and 6 introduces a PTC[9]. The splicing alterations can be caused by mutations within the *CD19* gene or by changes in the expression of *trans*-acting RBPs. For instance, it has been shown that the known splicing regulator SRSF3 binds to *cis*-regulatory elements within *CD19* exon 2 to promote its inclusion[9]. Of note, alternative *CD19* isoforms showing exon 2 skipping were observed to pre-exist in patients prior to CART-19 therapy[16], suggesting that *CD19* splicing patterns may harbour predictive information and could be modulated to re-establish sensitivity to CART-19-mediated killing. However, Orlando and co-workers suggested that alternative splicing changes in B-ALL patients are present in diagnostic samples already (albeit at low frequencies) and may not contribute meaningfully to CD19 epitope loss[5]. We, therefore, set out to investigate *CD19* alternative splicing and its molecular determinants in B-ALL in more detail.

High-throughput mutagenesis screens combined with next-generation sequencing provide comprehensive insights into the regulatory code of splicing[19–22]. The interpretation of such data is challenging, as the mutation effects often depend on other mutations and are typically most pronounced at intermediate exon inclusion levels[19,20,23]. We and others have shown by mathematical modelling that kinetic models account for the context-dependence of mutation effects on splice isoforms[19,20]. By utilising these models, systems-level insights can be gained into complex *cis*-regulatory landscapes, effects of *trans*-acting RBPs, and principles of splicing regulation[19,20,24].

In this work, we combine B-ALL patient data with high-throughput mutagenesis, mathematical modelling and RBP knockdowns to comprehensively characterise *cis*-regulatory mutations and *trans*-acting RBPs controlling *CD19* exon 2 splicing. Unlike previous mutagenesis screens, we determine all intronic and exonic mutation effects in a 1.2 kb region and quantify the abundance of 100 alternative isoforms, including intron 2 retention and alternative 3′/5′ splice site usage. Many of these isoforms encode for a non-functional CD19 protein and are therefore likely to impair CART-19 therapy. By in silico analyses and RBP knockdowns, we identify *trans*-regulators of *CD19* splicing that promote the production of the therapy-impacting isoforms. Taken together, our dataset allows for a systems-level understanding of the splicing code and provides a comprehensive resource of predictive markers for CART-19 therapy resistance.

Results

CART-19 patients show increased *CD19* intron 2 retention after relapse. To resolve the contribution of *CD19* splicing in CART-19 therapy, we re-analysed RNA-seq data from Orlando and co-workers[5], in which B-ALL cells of 17 patients were sequenced at initial screening and after relapse. In contrast to the original study, we expanded the analyses to intron retention events surrounding *CD19* exon 2. We found that the average frequency of intron 2 retention across patients is unexpectedly high (63%) before therapy and significantly increases to 82% after relapse (P value = 0.009, Wilcoxon signed-rank test; Fig. 1a, b). The trend towards higher intron 2 retention in the therapy-resistant tumours is preserved in seven out of nine individual patients that were sequenced both before therapy and after relapse (Fig. 1b). Since the resulting isoform does not encode a functional CD19 protein, this suggests that increased intron 2 retention contributes to CART-19 therapy resistance, as reported in a recent study[11].

Given the high prevalence of intron 2 retention even before CART-19 therapy, we extended our analysis to 220 B-ALL patients from the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) programme. Although these patients had not been treated with CART-19, intron 2 retention appears as the predominant isoform in almost all of them (Fig. 1c, Supplementary Fig. 1a, b). This suggests that the cancer cells generally exhibit *CD19* mis-splicing. Interestingly, strong intron 2 retention is also observed in immature B-cell precursors from healthy donors, whereas it is negligible in mature B cells (Fig. 1d). Therefore, incomplete B-cell differentiation in B-ALL may be accompanied by *CD19* mis-splicing which is further aggravated during CART-19 therapy.

Somatic mutations in relapsed patients cause splicing alterations. To learn about the genetic causes of the splicing alterations during relapse, we took a closer look at the mutations accumulating in the B-ALL patients of the Orlando study[5]. The majority of relapsed patients (12 out of 17) harbour somatic mutations within the *CD19* gene, including frameshift insertions, deletions and single nucleotide missense variants. We selected nine mutations in exons 2 or 3 from eight patients for further analysis (Supplementary Table 1). To test for effects on splicing, we constructed a minigene reporter that harbours *CD19* exon 1–3 including the two intervening introns 1 and 2 (Fig. 1e). We confirmed that the minigene gives rise to the same transcript

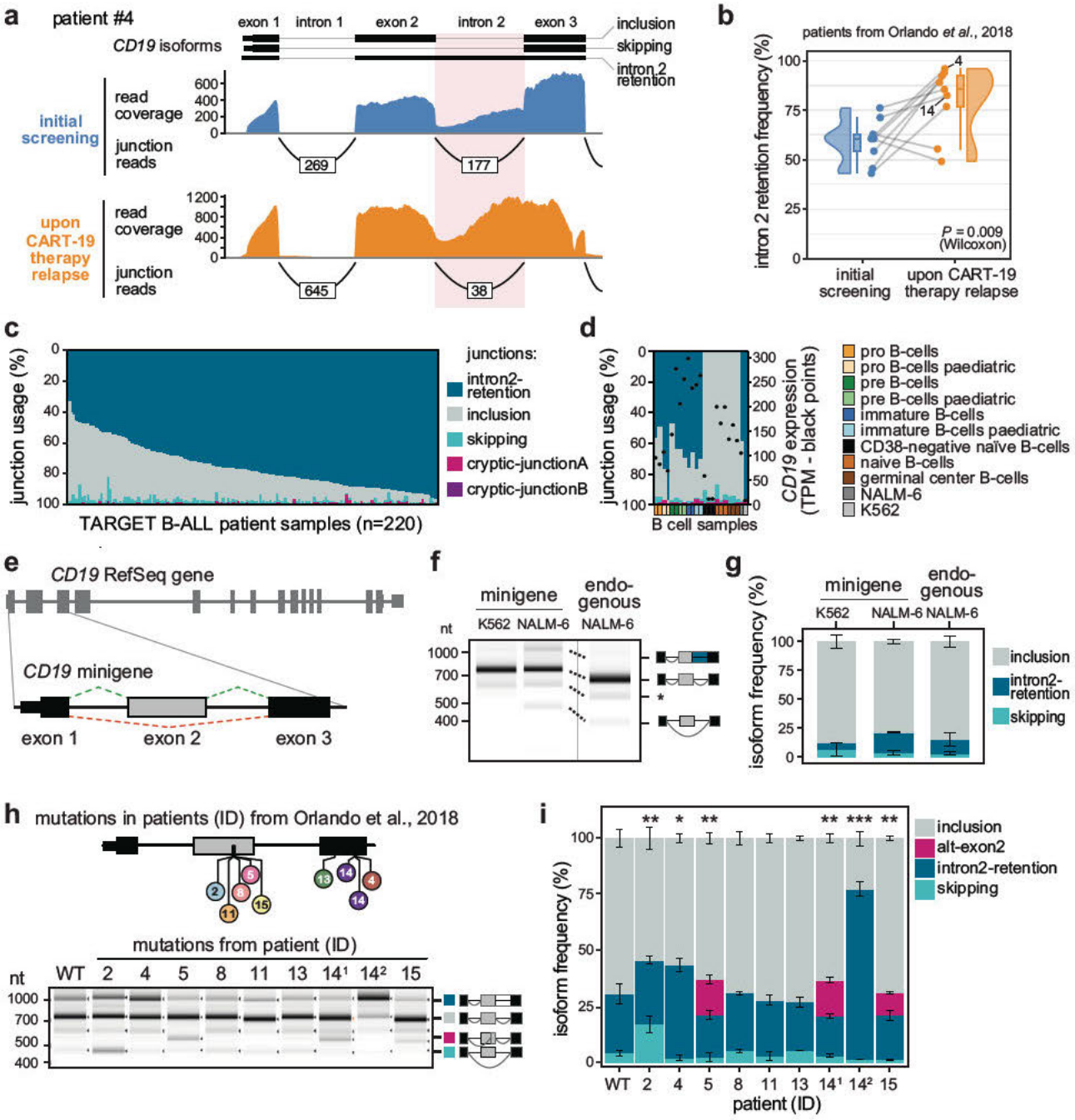


Fig. 1 Mutations from B-ALL patients cause *CD19* mis-splicing. a Patient #4 shows increased *CD19* intron 2 retention after CART-19 therapy relapse. Re-analysed RNA-seq data from Orlando et al.[5]. Selected isoforms (GENCODE) are shown. **b** Intron 2 retention increases in B-ALL patients after CART-19 therapy relapse. Intron 2 retention frequency (as % of all isoforms) is shown for nine patients with matched RNA-seq data at screening and after relapse. *P* value = 0.009, one-sided paired Wilcoxon signed-rank test. Grey lines connect matched samples. Boxes represent quartiles, centre lines denote 50th percentile, and whiskers extend to most extreme values within 1.5× interquartile range (IQR). **c, d** Intron2-retention is the predominant isoform in B-ALL patients and pre B cells. Stacked bar chart shows relative usage (percent selected index, PSI; left *y* axis) of junctions originating from exon 3 (Supplementary Fig. 1a) in 220 patients from the TARGET B-ALL programme (**c**) and normal B cells[44, 75] ($n = 21$) (**d**). Black dots in **d** indicate total *CD19* mRNA expression, in transcripts per million (TPM; right *y* axis). Cell lines NALM-6 and K562 are shown for comparison. **e** The *CD19* minigene spans exons 1–3 and the intervening introns from the *CD19* gene. **f, g** The minigene generates the same isoforms as the endogenous *CD19* gene in NALM-6 cells. Gel-like representation (**f**) and quantification (**g**) of semiquantitative RT-PCR showing isoforms intron2-retention (blue), inclusion (grey) and skipping (turquoise) for the WT minigene in NALM-6 cells. Isoforms of *CD19* gene in NALM-6 cells are shown for comparison. Asterisk indicates a previously reported RT-PCR artefact[66] (see Methods). Error bars indicate standard deviation of mean (s.d.m.), $n = 3$ replicates. *P* value > 0.1 for all isoforms, one-way ANOVA. **h, i** Patient mutations cause splicing changes in the *CD19* minigene. Top: Location of the tested mutations. Patient IDs as reported in Orlando et al.[5]. 14.1 and 14.2 correspond to distinct mutations from patient #14. Gel-like representation (**h**) and quantification (**i**) of semiquantitative RT-PCR as in **f, g**. Additional isoform alt-exon2 (purple) includes a truncated version of exon 2. Error bars indicate s.d.m., $n = 3$ replicates. **P* value < 0.05, ***P* value < 0.01, ****P* value < 0.001, two-sided Student's *t* test. Source data including *P* values are provided as a Source Data file.

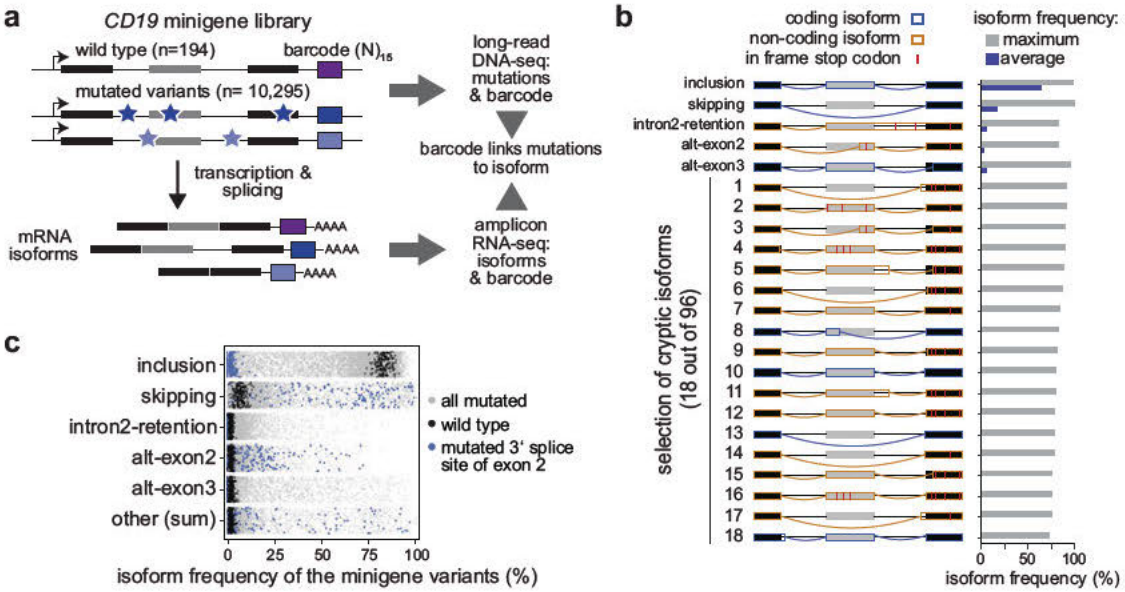


Fig. 2 High-throughput mutagenesis identifies splicing-affecting mutations and cryptic isoforms in the *CD19* minigene. a High-throughput detection of splicing-affecting mutations and cryptic isoforms. Mutagenic PCR creates mutated minigene variants (top) that upon transfection into NALM-6 cells give rise to alternatively spliced transcripts (bottom). Mutations (stars) and corresponding splicing products are characterised by DNA and RNA sequencing, respectively, and linked by a unique 15-nt barcode sequence in each minigene (coloured boxes). Black and grey boxes depict constitutive and alternative exons, respectively. **b** A large number of *CD19* splice isoforms arise in the minigene library. *CD19* splice isoforms with the highest maximal isoform frequency across all 9321 minigene variants. Schematic representation (left) of 5 major and 18 cryptic isoforms depicts exons 1–3 (boxes) and introns (horizontal lines) with splice junctions for each isoform (arches). Colour indicates coding potential (blue, coding; orange, non-coding). In-frame stop codons are indicated by red lines. Bar graph (right) shows average and maximal isoform frequency across all minigenes. Cryptic isoforms are sorted by maximal isoform frequency (Supplementary Data 2). **c** Inclusion isoform dominates in WT minigenes, whereas mutated variants show broad spread in all major isoforms. Frequencies of five major isoforms in replicate 1 for all wild type (black; $n = 195$) and mutated (grey; $n = 9476$) minigenes in the library. Minigene variants harbouring a mutation in the 3′ splice site of exon 2 ($n = 174$) are highlighted in blue. "Other" refers to the sum of 96 cryptic isoforms.

isoforms with quantitatively similar frequencies as the endogenous gene in the human B-ALL cell line NALM-6 (Fig. 1f, g, Supplementary Fig. 1c).

When introducing the patient mutations into our minigene reporter, we found that six out of nine tested mutations lead to the production of alternative *CD19* isoforms linked to CART-19 therapy resistance (Fig. 1h, i, Supplementary Fig. 1d): The mutation from patient #2 induces exon 2 skipping, while mutations from patients #4 and #14 (#14.2) cause intron 2 retention. The latter mirrors the increase of this isoform in the same patients after CART-19 therapy relapse (Fig. 1b). In addition, three mutations enhance the production of an additional isoform that uses an alternative 3′ splice site in exon 2 (termed alt-exon2; mutations from patients #5, #14.1 and #15). We note that as reported by Orlando and co-workers[5], most of the patient mutations also introduce frameshifts and therefore perturb CD19 protein expression by disrupting the open reading frame. Interestingly, splicing effects and frameshifts potentially interact in a non-intuitive way: For instance, the deletion in patient #5 causes a frameshift, but at the same time activates an alternative 3′ splice site (alt-exon2) which restores the open reading frame (Supplementary Fig. 1e). Thus, taking splicing information into account is essential to understand whether a targetable CD19 protein is generated in a patient harbouring *CD19* mutations.

High-throughput screening of alternative splicing of *CD19* exons 1–3. To systematically study the effects of point mutations on *CD19* exons 1–3 splicing, we adopted our previously developed massively parallel splicing reporter assay[19] (Fig. 2a). To this end, we randomly introduced point mutations as well as short insertions and deletions into the *CD19* minigene reporter by error-prone PCR. This yielded a pool of 10,295 minigene variants, each with a different set of mutations and tagged with a unique 15-nt barcode sequence. As an internal control, 194 wild type (WT) minigenes with distinct barcodes were added. Mutations in all minigene variants were mapped using targeted long-read DNA sequencing (DNA-seq, PacBio SMRT-seq, Supplementary Fig. 2a, b) and validated for 30 minigene clones via Sanger sequencing. The DNA-seq data shows that the minigene variants contain on average 9.7 mutations (Supplementary Fig. 2c). This allows for a comprehensive characterisation of the mutation landscape, as each position is on average mutated in 80 different minigene variants and 90% of the mutations are present in at least four distinct minigene variants (Supplementary Fig. 2d, e). To measure splicing outcomes, the minigene pool was transfected into NALM-6 cells and the resulting transcripts were quantified by targeted RNA sequencing (RNA-seq) using 350 nt + 250 nt paired-end reads (Illumina MiSeq, Supplementary Figs. 2a, 3a). We detected around 100 different splice isoforms (see below) which were unambiguously identified by paired-end sequencing. Two replicate experiments showed a high correlation in the measured isoform frequencies (R between 0.91 and 0.98 for the different isoforms, Supplementary Fig. 3b). Based on the common barcode sequence, information from DNA and RNA sequencing could be combined, linking mutations at the DNA level to frequencies of RNA splice isoforms for a total of 10,295 minigenes in two replicate experiments (Supplementary Data 1).

Therapy-impacting isoforms accumulate in response to numerous point mutations. To our surprise, the screen revealed a high complexity of *CD19* exon 1–3 splicing, with a total of 101

alternative isoforms occurring with a frequency of ≥5% of all transcripts in at least two minigene variants (Supplementary Data 2). Out of these, the five major isoforms exceed 1% in WT minigenes, whereas the others, termed cryptic isoforms, only accumulate in mutated minigene variants (Fig. 2b). In WT, the most abundant major isoform by far is exon 2 inclusion (termed "inclusion"), followed by exon 2 skipping (termed "skipping") and intron 2 retention (termed "intron2-retention"). Two additional major isoforms in WT originate from alternative 3′ splice site usage within exon 2 (alt-exon2) and 3 (alt-exon3) (Fig. 2b, c). Notably, alt-exon2 uses the same splice junction that we had observed upon introducing patient mutations into the *CD19* minigene (Fig. 1i). As expected, the measured frequencies for the major isoforms show little variance for the 194 unmutated WT minigenes (standard deviation <6%, Fig. 2c). In contrast, many mutated minigene variants show strong changes relative to WT, suggesting a large impact of specific mutations on splicing outcomes (Fig. 2c). For instance, all minigenes with a mutation in the 3′ splice site of exon 2 lose the inclusion isoform, accompanied by strong alterations in the remaining major isoforms. Taken together, these observations support the accuracy of our screening results.

All major isoforms, except exon 2 inclusion, could contribute to therapy resistance either by generating an altered CD19 protein lacking a functional CART-19 epitope or by decreasing its production. Our unbiased screening approach extends the list of potentially therapy-impacting *CD19* mutations, since 1721 out of 9127 mutated minigenes show exon 2 skipping, intron 2 retention and/or alt-exon2 isoform frequencies of >25% (Fig. 2c). However, since the minigene variants carry on average 9.7 point mutations, the observed splicing changes represent the combined effects of several mutations. To extract the impact of individual mutations, we adapted our previous mathematical modelling framework[19] and implemented a multinomial logistic regression approach. Here, the splicing change in each minigene variant is described as the sum of the underlying point mutation effects (Fig. 3a, see Methods). These single-mutation effects are unknown and are determined by simultaneously fitting the model to all minigene measurements. Thereby, we were able to infer the individual effects of 4255 point mutations on the five major isoforms (Fig. 3a, Supplementary Fig. 4a). We validated the reliability of this model in describing combined mutations using a 10-fold cross-validation approach, in which we left out 10% of all minigene variants from fitting and were able to accurately predict them after model fitting (Pearson correlation coefficients 0.68–0.95; Fig. 3b, Supplementary Fig. 4b). In particular, isoforms abundant in WT or strongly accumulating in response to a large number of mutations (inclusion, skipping and alt-exon3) were predicted with high accuracy, whereas the prediction power was slightly lower for isoforms with a worse signal-to-noise ratio (intron2-retention and alt-exon2; Fig. 2c). Furthermore, we estimated that the model performed well in predicting single-mutation effects, as soon as a mutation occurred in three or more minigenes in the dataset (Supplementary Fig. 4c), which applied to 90% of all mutations (Supplementary Fig. 2e).

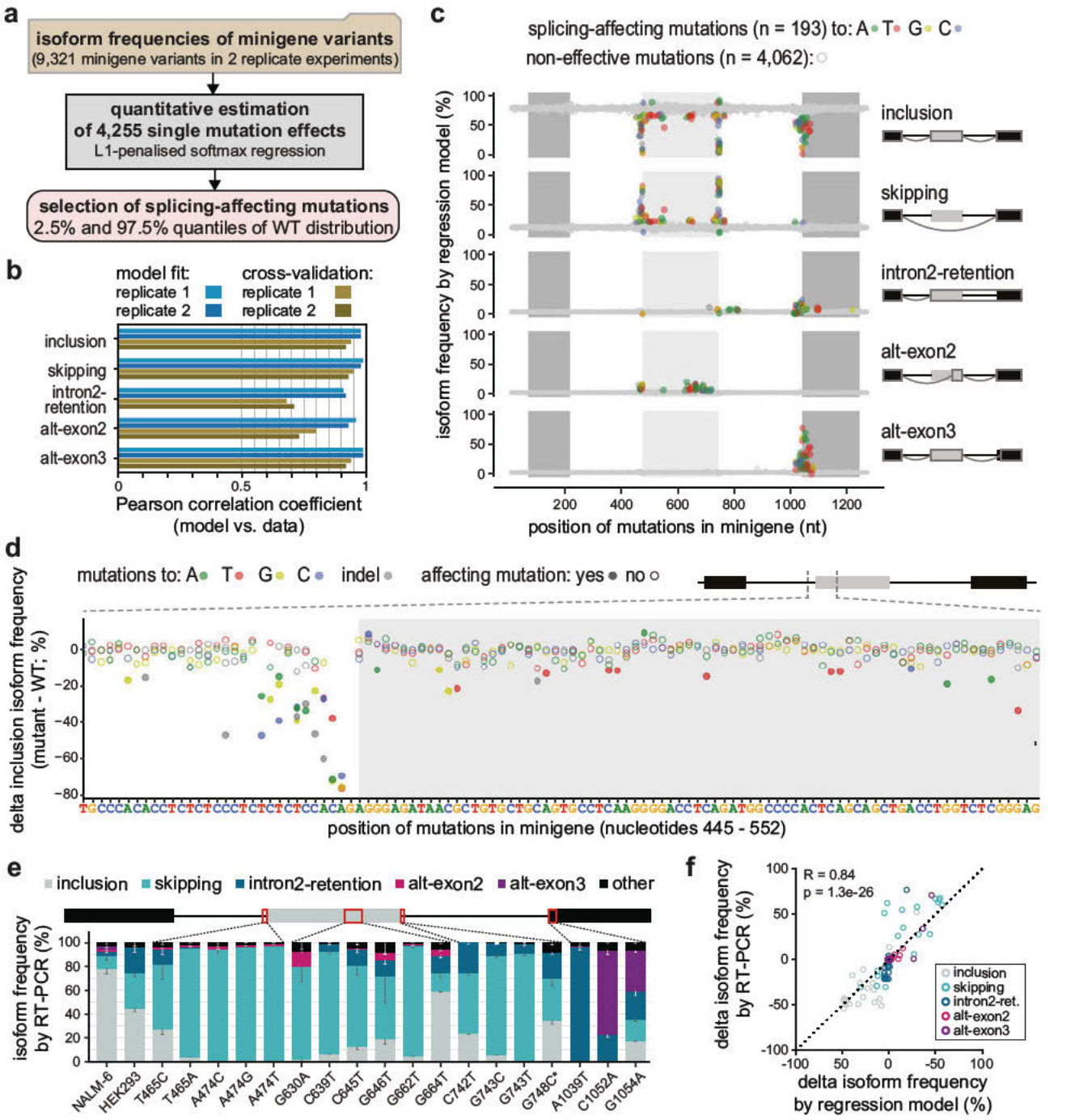
Out of 4255 quantified single-mutation effects, we find 193 splicing-affecting mutations that significantly alter the frequency of at least one isoform in the two replicates beyond the 2.5 and 97.5% quantiles of the WT minigene distribution (Fig. 3c, Supplementary Data 3, Source Data file). 37 of these splicing-affecting mutations overlap with single nucleotide variants (SNVs) that were previously reported in the human population from whole-genome or exome sequencing data, as well as reported cancer-associated mutations, with some of them predicted to have a pathogenic or likely pathogenic effect in the disease context (Supplementary Data 3). The strongest mutation effects accumulate around the four main splice sites and throughout exon 2 and correspond to the core *cis*-regulatory elements, such as splice site dinucleotides, branchpoint and polypyrimidine tract, as well as auxiliary elements (Fig. 3c, d). In particular, 21% of all positions within exon 2 (55 out of 267 nt) harbour at least one splicing-affecting mutation for any isoform, suggesting that *CD19* exon 2 is densely packed with *cis*-regulatory elements.

Inspecting in more detail the 83 mutations that specifically impact *CD19* exon 2 skipping, we find them to cluster within and around exon 2 (odds ratio 2.06 for such mutations to occur inside exon 2, P value = 0.002614, Fisher's exact test). In addition, we observe smaller clusters of mutations within the introns and flanking constitutive exons which likely represent more distal *cis*-regulatory elements (Fig. 3c). Similarly, we explored the 54 splicing-affecting mutations impacting on intron2-retention. As expected, strongest effects are observed at the splice sites of intron 2. In addition, we find clusters of splicing-affecting mutations in intron 2 and exon 3 that might reflect important *cis*-regulatory elements. The predicted effects of all mutations on the five major isoforms can be explored in the associated Source Data file.

To test a subset of the regression predictions, we generated 19 minigenes with individual point mutations that are predicted to affect at least one isoform, including two previously reported SNVs (Supplementary Fig. 5a, Supplementary Data 4). Using semiquantitative RT-PCR, we were able to confirm that mutations near the splice sites of exon 2 predominantly lead to exon 2 skipping, whereas mutations in exon 3 result in intron2-retention and/or alt-exon3 formation (Fig. 3e, Supplementary Fig. 5b). Overall, the splicing measurements for the individual minigenes show high correlation with the regression predictions for the respective mutations across all five major isoforms (Fig. 3f, Supplementary Fig. 5c). In conclusion, our combined screening and modelling approach quantitatively describes alternative splicing of *CD19* exons 1–3 by predicting the effects of all individual point mutations and combinations thereof. Our screen thereby represents a comprehensive resource for the identification of mutations with potential clinical relevance in CART-19 therapy resistance.

Cryptic isoforms destroy the *CD19* ORF and are associated with recurrent mutations. Besides the five major isoforms, the *CD19* exons 1–3 can give rise to 96 cryptic isoforms which are rare (<1%) in WT, but accumulate upon certain mutations (Fig. 2b, Supplementary Data 2). The cryptic isoforms involve a total of 71 cryptic splice sites (Fig. 4a). Of note, 33 of these cryptic isoforms make up >50% of total transcripts and are therefore dominant in certain minigene variants (Fig. 2b, c). To assess whether these cryptic isoforms impact on CD19 epitope presentation, we analysed their coding potential and found that the vast majority of cryptic *CD19* isoforms (78 out of 96) show a frameshift and/or carry a PTC (Fig. 4b). This will either lead to the production of truncated CD19 peptides that likely do not allow for presentation on the cell surface[15] or will induce nonsense-mediated mRNA decay of the cryptic isoforms and will hence reduce *CD19* transcript and protein levels.

To derive a mechanistic understanding of cryptic isoform biogenesis, we analysed the underlying point mutations. To this end, we calculated a prevalence score which quantifies the degree of association between an isoform and a point mutation by multiplying: (i) the frequency of a mutation being present if the isoform level is high (>5%), and (ii) the frequency of the isoform level being high given that the mutation is present. A prevalence score of 1 indicates perfect correspondence between mutation and isoform, whereas a prevalence score of 0 is observed if they are unrelated. This score-



based analysis revealed 38 mutation-isoform pairs with prevalence scores > 0.25 which could explain the genesis of 36 cryptic isoforms based on 31 point mutations (Supplementary Fig. 6a, Supplementary Data 2). The remaining 60 cryptic isoforms do not show a specific association, implying that they can either be generated by multiple redundant mutations, or that our screen lacks sufficient coverage to support a reliable association. To directly test the predicted associations, we introduced five mutations with a specific association to a cryptic isoform in our minigene reporter (C535G, chr16: 28932405, prevalence score = 0.18; C806A, chr16:28932676, 0.68; A827T, chr16:28932697, 0.93; C864G, chr16:28932875, 1; G1005A, chr16:28932734, 0.89). Semiquantitative RT-PCR confirmed that all five tested mutations lead to the appearance of the associated cryptic isoform (Fig. 4c, d).

Altogether, our analysis provides a list of 31 mutations that are likely to trigger cryptic isoform formation. Importantly, the resulting cryptic isoforms show a maximum usage of up to 91% (Supplementary Data 2), which is expected to drastically interfere with normal *CD19* splicing, protein production, and subsequent epitope presentation. Screening for the occurrence of the 96 cryptic isoforms in the TARGET B-ALL patient samples, we readily detected two junctions of cryptic isoforms that had been present already prior to CART-19 therapy (Fig. 1c, Supplementary Fig. 1a). Other cryptic isoforms predicted from our screen were not found in these patients that had not been treated with CART-19 therapy, but could already exist subclonally and/or may only emerge under the selective pressures of CD19-directed immunotherapy. The same applies to the associated mutations

Fig. 3 Quantitative modelling predicts single-mutation effects on splice isoforms. a Based on the experimentally measured frequencies of five major isoforms in 9321 minigene variants (top box), a softmax regression model was formulated to estimate 4255 single-mutation effects (middle box) using L1 penalisation. Splicing-affecting mutations were selected for each isoform based on their respective empirical WT frequency distribution using the 2.5% and 97.5% quantiles as cutoff. **b** The model performs well in fitting and 10-fold cross-validation. Bars show Pearson correlation coefficients between model and data for two replicates and each of the five isoforms across all combined mutation minigenes considered in model training and validation, respectively (Supplementary Fig. 4a, b). **c** Splicing-affecting mutations accumulate in distinct regions around exons 2 and 3. Landscape of model-predicted single-mutation effects on five major isoforms. Predicted isoform frequencies are plotted as a function of the position of a mutation. Colours indicate nucleotide substitution of splicing-affecting point mutations (see legend), and non-effective mutations (grey). **d** Zoom-in shows model-predicted delta inclusion isoform frequency (frequency for a point mutation - frequency in WT) for nucleotides 445–552 of the minigene. Splicing-affecting mutations are highlighted as filled circles. **e** Model validation by splicing analysis of 19 minigene variants containing single point mutations. Isoform frequencies (in %) of the five major isoforms (see legend) are shown as mean values of three biological replicates (error bars, s.d.m.). 'NALM-6', splicing pattern of WT minigenes (RNA-seq) in the mutagenesis screen, 'HEK293', RT-PCR-based quantification of the baseline minigene containing mutation G742C (see Methods) in HEK293 cells. G748C* is a minigene containing G748C but lacking G742C. Schematic representation of *CD19* minigene (top) highlighting mutated regions (red rectangles). Error bar represent s.d.m., $n = 3$ replicates. **f** Splicing outcomes from **e** (*y* axis) are related to single-mutation predictions of the regression model (*x* axis; mean of two fits, each explaining one mutagenesis replicate). Changes in isoform frequency of the major isoforms (see legend) are expressed as differences (delta) relative to the baseline. Pearson correlation coefficient and *P* value (two-sided) were calculated over all isoforms (see Supplementary Fig. 5c for correlations of individual isoforms). Source data are provided as a Source Data file.

identified from our screen which were also not present in the TARGET B-ALL data (Supplementary Data 4).

The cryptic isoforms are caused by mutations that disrupt or create splice sites. Due to their potential clinical relevance, we wanted to learn more about how the mutations activate the cryptic isoforms. We found that the majority of mutations with a prevalence score > 0.25 are either in close proximity or directly overlap with the associated cryptic splice site (77.4% with distance <5 nt; odds ratio 7.55, *P* value = 1.793e-07, Fisher's exact test; Fig. 4e). Further inspection showed that the underlying mutations either destroy the original splice site (7.9%) or generate a new cryptic splice site (57.9%). Hence, the cryptic isoforms do originate from the generation or destruction of core *cis*-regulatory elements rather than affecting auxiliary elements.

Currently, major efforts are ongoing to implement artificial intelligence (AI) tools to predict the effect of clinical variants on the splicing outcome. We therefore tested whether the state-of-the-art neural network SpliceAI[25], which predicts changes in the splicing patterns induced by single point mutations, captures the gain and loss of splice sites in *CD19*. We applied SpliceAI using all possible single-point mutations in the *CD19* minigene as an input. Similar to the results from our mutagenesis screen (Fig. 4a), SpliceAI predicts cryptic splice site activation by mutations throughout the minigene, with an increased density around the 3′ splice site of exon 3 (Supplementary Fig. 6b). All SpliceAI-predicted mutations are close to the affected cryptic splice sites (Supplementary Fig. 6c). Hence, SpliceAI successfully reflects the global landscape of mutation-induced cryptic splice site activation in the *CD19* minigene.

With respect to the accuracy of the individual predictions, we found that 10 out of 38 mutations with strong SpliceAI predictions (SpliceAI score > 0.5) indeed lead to the accumulation of splice isoforms with the corresponding cryptic splice sites in the experimental data (prevalence score > 0.25, Fig. 4f). In the remaining 28 cases, either weak overall cryptic splice site activation occurred in the data (9 cases) or a different cryptic splice site was activated than predicted by SpliceAI (19 cases; Supplementary Fig. 6b). In quantitative terms, the likelihood of a cryptic splice site activation according to the SpliceAI prediction ("SpliceAI score") is correlated to the magnitude of the prevalence score linking the mutation to the corresponding cryptic isoform in our screen (Fig. 4f). Overall, the comparison supports that SpliceAI can guide the interpretation of mutation effects in clinical samples, though direct experimental validation is necessary. Due to the robust performance of SpliceAI, we decided to predict splice-changing mutations throughout the entire *CD19* gene and overlapped them with publicly reported single nucleotide variants (SNVs; Supplementary Fig. 6c). These predictions and variant overlap are provided as a resource (Supplementary Data 5) and can be used to evaluate the impact of new patient mutations on *CD19* splicing in the future.

From our mutagenesis data, we found that the cryptic isoforms arise from numerous 3′ and 5′ cryptic splice sites that distribute over the entire minigene and accumulate at exon 3 (Fig. 4a). In line with their high prevalence, 26 cryptic splice sites reach >50% usage upon certain mutations, particularly around the start of exon 3. We hypothesised that cryptic splice site activation occurs in exon 3 because its canonical splice site can be outcompeted by neighbouring cryptic sites. To test this, we scored the strength of local consensus sequences using MaxEntScan[26], and indeed found that the 3′ splice site of exon 3 is weak compared to all other canonical splice sites of *CD19* exons 1–3 (Fig. 4g, Supplementary Fig. 6e, f). In line with our hypothesis, mutations around the 3′ splice site of exon 3 frequently create stronger splice sites than elsewhere in the minigene that exceed the strength of the canonical 3′ splice site of exon 3 (Fig. 4g). This suggests that weak splice sites are particularly vulnerable to the activation of competing cryptic splice sites and should be of particular interest when assessing the impact of clinical variants on splicing outcomes.

An extensive network of RBP regulators might drive *CD19* mis-splicing. Besides *CD19* mutations, CART-19 therapy resistance may also stem from altered expression of *trans*-acting RBPs which bind to the *CD19* pre-mRNA to control alternative splicing. To identify putative RBP regulators, we explored publicly available databases containing experimentally determined RBP binding motifs (ATtRACT[27] and oRNAment[28]). Furthermore, we included RBP binding information from the public resource of ENCODE eCLIP datasets[29]. Since the *CD19* mRNA is hardly expressed in the ENCODE cell lines and binding events in *CD19* can therefore not be directly extracted, we employed the prediction algorithm DeepRiPe[30]. The underlying neural network has been trained on the PAR-CLIP and ENCODE eCLIP datasets and thereby allows to predict changes of RBP binding upon mutation in any RNA sequence. In combination, these tools predict a total of 198 RBPs to bind within *CD19* exons 1–3 (ATtRACT: 62 RBPs; oRNAment: 70 RBPs) or to decrease (or increase) binding upon mutation (DeepRiPe: 128 RBPs; Fig. 5a, b, Supplementary Fig. 7a).

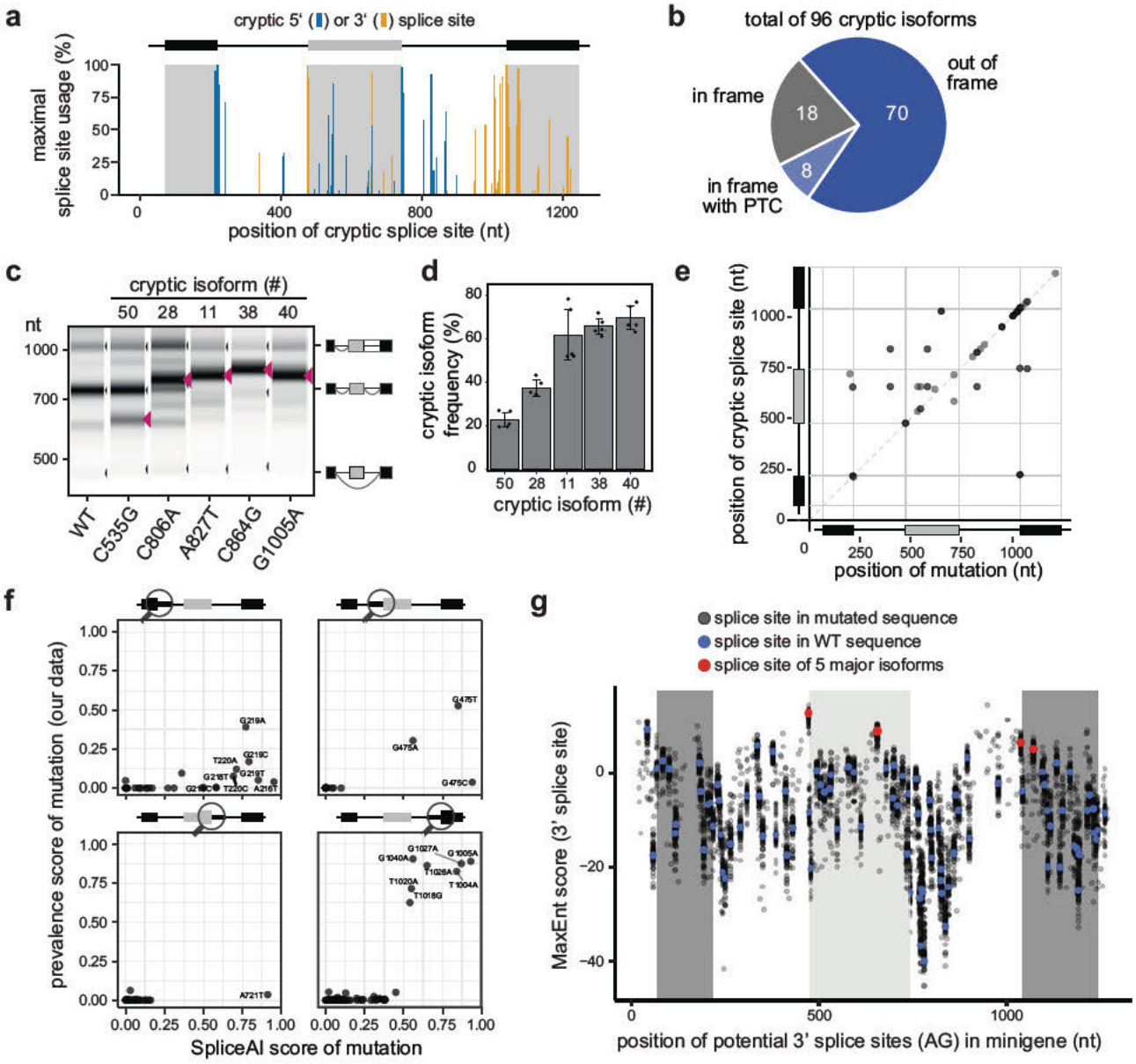


Fig. 4 *CD19* mutations frequently activate cryptic splice sites. a Alternative splicing of *CD19* minigene variants involves 71 cryptic splice sites. Splice site usage (sum of junction-spanning reads in a splice site / total number of reads in minigene) was calculated for each minigene variant. Maximum usage across all minigenes is plotted against the corresponding position to the cryptic splice sites. **b** Cryptic isoforms code for non-functional CD19 proteins. Out of 96 cryptic isoforms, 8 run into a premature termination codon (PTC) and 70 are out-of-frame. The remaining 18 remain in frame, but are shortened or extended relative to the reference inclusion isoform. **c**, **d** Experimental validation of five point mutations associated with distinct cryptic isoforms. Predicted cryptic isoforms are indicated by red arrowheads. Gel-like representation (**c**), with major isoforms indicated on the right, and RT-PCR quantification (**d**). Error bars indicate s.d.m., $n = 3$ replicates. **e** Mutations leading to cryptic isoforms are often located within or near cryptic splice sites. For 31 cryptic isoforms that are highly associated with a mutation (prevalence score > 0.25; *y* axis), the position of the mutation (*x* axis) was related to the used cryptic splice site (*y* axis). **f** SpliceAI correctly predicts single mutations leading to the generation of cryptic isoforms. SpliceAI scores of 0–1 reflecting the probability to gain a cryptic splice site in response to a mutation (see Methods). Scatterplots compare the SpliceAI score against the prevalence score (association of a mutation with a cryptic isoform) from our data, for 254 mutation-splice site pairs that match in their positions with SpliceAI. Separate panels are shown for each canonical splice site (circle in schematic minigene representation). **g** Exon 3 harbours a weak 3′ splice site and is preceded by many potentially competing cryptic 3′ splice sites. Dotplot shows splice site strengths (MaxEnt score) for putative 3′ splice sites (AG dinucleotides) in the *CD19* minigenes. MaxEnt score was calculated in 23-nt sliding window for WT sequence (red and blue dots) and hypothetical mutant minigenes with all possible single-point mutations (grey dots). 3′ splice sites used in the five major isoforms are highlighted in red. Source data are provided as a Source Data file.

To link the putative RBP regulators to the observed splicing changes, we overlaid the predicted binding sites (or predicted mutations for DeepRiPe) with splicing-affecting mutations from our screen. Overall, we find that 79% and 60% of ATtRACT and oRNAment binding sites, respectively, overlap with a splicing-affecting mutation (affecting any of the five major isoforms). Furthermore, 105 (5%) of the mutations predicted to change RBP binding by DeepRiPe overlap with splicing-affecting mutations, suggesting that modulating RBP binding at these sites may have a functional impact on *CD19* splicing (Fig. 5c, Supplementary

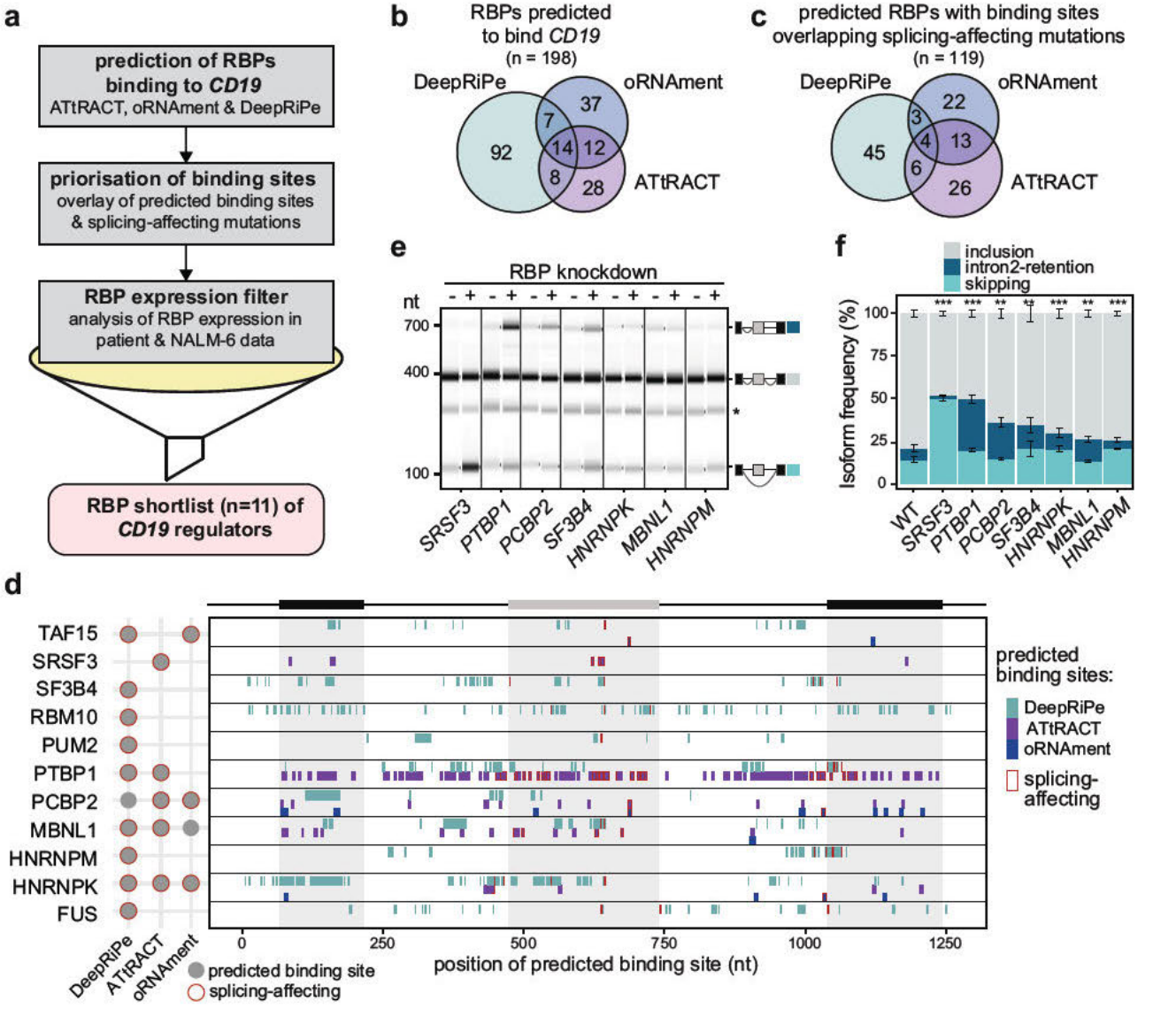


Fig. 5 In silico predictions identify RBP regulators of *CD19* alternative splicing. a Pipeline for the identification of potential RBP regulators of *CD19* splicing. Starting with in silico predictions, we obtained 198 candidate RBPs with predicted binding motifs (ATtRACT/oRNAment) or predicted differential binding upon mutation (DeepRiPe). These were prioritised by overlapping with the splicing-affecting mutations from our screen. Additionally, based on publicly available RNA-seq data, we required a minimum mean expression in RNA-seq data from B-ALL patients[31] and NALM-6 cells[75]. Together with literature information, we shortlisted 11 candidate RBPs for knockdown (KD) experiments, including SRSF3 as a positive control. **b**, **c** In silico analyses predict dozens of RBPs binding to *CD19*. Venn diagrams show overlap of RBPs in initial predictions (**b**) and after overlay with splicing-affecting mutations (**c**). **d** The 11 candidate RBPs are predicted to bind throughout the *CD19* minigene region. For each RBP, the binding sites predicted by ATtRACT and oRNAment and disrupting mutations predicted by DeepRiPe are indicated (see legend). Sites overlapping with splicing-affecting mutations are framed in red. The schematic summary (left) shows that all 11 candidate RBPs have at least one predicted site that overlaps with a splicing-affecting mutation. A full list of predicted binding sites (ATtRACT/oRNAment) and differential binding mutations (DeepRiPe) is provided in Supplementary Data 6. **e**, **f** Seven RBP KDs significantly change *CD19* splicing. Gel-like representation (**e**) and quantification (**f**) of semiquantitative RT-PCR showing detected isoforms exon 2 inclusion (grey), intron 2 retention (blue), and skipping (turquoise) from the endogenous *CD19* gene in KD and control NALM-6 cells. Asterisk indicates a previously reported RT-PCR artefact[66] (see methods). Error bars indicate s.d.m., $n = 3$ replicates. $^{**}P$ value < 0.01, $^{***}P$ value < 0.001, two-sided Student's *t* test. Measurements for all 11 KD experiments are shown in Supplementary Fig. 8c, d. Source data including *P* values are provided as a Source Data file.

Fig. 7a). By merging these sets, we obtained a list of 119 RBPs that may regulate splicing by binding to *CD19* exons 1–3 (Supplementary Data 6). Most of these are expressed in cancerous B cells from B-ALL patients from[31] (80 with mean FPKM [fragments per kilobase of transcript per million mapped reads] > 10; Supplementary Fig. 7b) and could thus modulate *CD19* splicing. Among these RBPs are SRSF3, a previously reported regulator of *CD19* splicing[9], but also new candidates such as PTBP1. Overall, the in silico predictions suggest the presence of an extensive RBP network that controls *CD19* splicing and may impact the CART-19 therapy success.

Depletion of PTBP1 and several other RBPs results in non-functional *CD19* isoforms. Based on our experimental data, in silico predictions, expression, literature information and manual curation, we shortlisted 11 RBP candidates for further analysis, including SRSF3 as a positive control (Fig. 5d). All 11 RBPs are

expressed in normal B cells and B-ALL patient samples from the TARGET B-ALL cohort (Supplementary Fig. 8a). To test their impact on endogenous *CD19* splicing, we generated NALM-6 cell lines stably expressing shRNAs against the shortlisted RBPs (depletion to <40% transcripts; Supplementary Fig. 8b). As previously described[9], knockdown of *SRSF3* leads to increased exon 2 skipping in the endogenous *CD19* transcripts, confirming that this SR protein is required for exon 2 inclusion (Fig. 5e, f). Importantly, we find that knockdown of six additional RBPs (*PTBP1*, *PCBP2*, *SF3B4*, *HNRNPK*, *MBNL1* and *HNRNPM)* has significant effects on *CD19* alternative splicing (Fig. 5e, f, Supplementary Fig. 8c, d). The knockdown of these factors reduces *CD19* exon 2 inclusion, while promoting intron2-retention and/or exon 2 skipping, thus shifting the cells towards expression of the relapse-associated *CD19* isoforms. This implies that reduced levels of these factors can impair targetable CD19 epitope expression.

PTBP1 stands out among the putative regulators as it shows the strongest effects on intron2-retention. This splicing event introduces a premature termination codon that likely reduces *CD19* transcript and protein expression via nonsense-mediated mRNA decay (Fig. 2b). In line with a role of PTBP1 in *CD19* mis-splicing in tumours, we find that patient samples from the TARGET B-ALL cohort on average show lower *PTBP1* mRNA expression compared to healthy B cells (Supplementary Fig. 8a). Within the B-ALL samples, *PTBP1* expression negatively correlates with *CD19* intron2-retention, as expected based on our knockdown experiments ($R = -0.24$; Fig. 6a, left). In addition, we investigated *PTBP2* mRNA expression, which is tightly repressed by the PTBP1 protein via alternative splicing and nonsense-mediated mRNA decay[32] and hence serves as a direct sensor for PTBP1 activity in the cells. Indeed, we find a strong correlation between increased *PTBP2* mRNA levels, i.e., lowered PTBP1 protein activity, and increased *CD19* intron2-retention ($R = 0.56$; Fig. 6a, right). To test for changes upon CART-19 relapse, we extracted *PTBP1* and *PTBP2* from expression data provided by the Orlando study[5]. Although we do not detect systematic changes in the *PTBP1* mRNAs levels, the *PTBP2* mRNA levels are significantly increased at relapse relative to screening, possibly indicating lowered PTBP1 protein levels (P value $= 0.037$, Wilcoxon rank-sum test; Fig. 6b). Together, these analyses suggest that PTBP1 is a regulator of *CD19* alternative splicing, which we decided to explore further.

PTBP1 recognises clusters of UC-rich motifs[33,34]. Remarkably, ATtRACT predicts almost 100 such PTBP1 binding motifs across the studied *CD19* region, including 25 that overlap with splicing-affecting mutations (Fig. 5d, Supplementary Data 6). Moreover, DeepRiPe predicts 78 mutations in 63 positions that change PTBP1 binding, out of which 10 are splicing-affecting in our screen (odds ratio 3.21, P value $= 0.002481$, Fisher's exact test). The high number of predicted binding sites suggests a partial redundancy, indicating that PTBP1 regulation might be difficult to disrupt with individual point mutations as introduced in our screen. To experimentally test if PTBP1 binds to the predicted sites, we performed PTBP1 iCLIP2 experiments[35] in NALM-6 cells. In line with a role in intron2-retention, we find extensive PTBP1 binding, particularly in intron 2, where it spreads over an extended cluster of predicted binding sites (Fig. 6c). This suggests that PTBP1 directly regulates *CD19* splicing via intron 2 binding.

Next, we chose to assess whether PTBP1-mediated splicing changes affect CD19 surface exposure on B cells. To test this, we performed siRNA-mediated knockdown of *PTBP1* in P493-6 and MHHCALL4 cells (Supplementary Fig. 9a, b) and confirmed that the knockdown increased levels of *CD19* intron2-retention in both cell lines (Supplementary Fig. 9c, d). Then, we measured CD19 protein surface expression using CD19 antibody staining and flow cytometry analysis (Supplementary Fig. 9e). Strikingly, we found that both cell lines show reduced CD19 surface exposure upon *PTBP1* depletion (Fig. 6d–f, Supplementary Fig. 9f). Thus, by interfering with CD19 protein expression on the cell surface, *PTBP1* depletion could indeed contribute to CART-19 therapy resistance.

Taken together, our data suggest that both *cis*-acting mutations and *trans*-acting RBPs can lead to unproductive *CD19* splicing which in turn disrupts CD19 epitope presentation. Therefore, the splicing-affecting mutations and RBP regulators identified in this work may harbour predictive information for CART-19 therapy success.

Discussion

Massively parallel reporter assays such as our high-throughput mutagenesis screen provide comprehensive insights into the regulatory code of splicing, as they characterise the complete set of *cis*-acting sequence mutations and reveal the binding sites of *trans*-acting RNA-binding proteins (e.g.,[19–21,36–38]). The interpretation of these datasets is challenging due to nonlinear interactions of individual mutation effects. For instance, competition effects in splicing reduce the impact of individual mutations at low and high isoform frequencies, i.e., depending on the mutational background[19,20]. In addition, other factors such as RBP expression patterns and cell type/tissue identity determine the effects of sequence mutations. Using kinetic modelling, we and others derived regression models taking competition in splicing into account, thereby showing that the effects of complex mutation combinations can be quantitatively described as the sum of individual mutation effects[19,20]. Thus, mutations seem to control splicing additively rather than synergistically, and this principle also holds for *CD19* splicing.

In our *CD19* mutagenesis dataset, we comprehensively characterise the full set of splice isoforms generated in response to thousands of sequence mutations. In particular, we find that cryptic splice site activation and thus alternative 3′ and 5′ splice site usage are common modes of alternative splicing. Intriguingly, such events do not require extensive sequence remodelling, but can often be triggered by single point mutations, as indicated by strong associations between putative cryptic isoforms and certain nucleotide substitutions. This suggests, in accordance with previous reports[39], that neighbouring splice sites frequently compete for spliceosome assembly, especially if the canonical splice site is comparably weak. While this finding shows the enormous isoform complexity that can arise already from such a simple exon configuration, it raises the question of how protein function can be robustly maintained, since most cryptic *CD19* splicing isoforms likely encode non-functional proteins.

Unlike previous mutagenesis screens which mainly focused on exonic sequence mutations, the present *CD19* dataset characterises the complete set of intronic and exonic mutations in a 1200 nt sequence stretch. The complete characterisation of *CD19* exons 1–3 required the use of long-read sequencing technology. Given that introns in human protein-coding genes on average span ~8.1 kb (GENCODE v31), the long-read sequencing methodology described in this work opens the approach for broad applications. For *CD19*, we find that strong mutation effects are mainly centred around canonical and cryptic splice sites, whereas in other examples such as *MST1R* exon 11 or *FAS* exon 6, mutation effects are more dispersed across intronic and exonic sequences[19,40]. This suggests that *CD19* exon 2 splicing may be controlled by multiple splicing enhancers that act redundantly and render inclusion less sensitive to individual point mutations[20]. Therefore, *CD19* exon 2 may require more specific perturbations and as we show here, does not only respond with

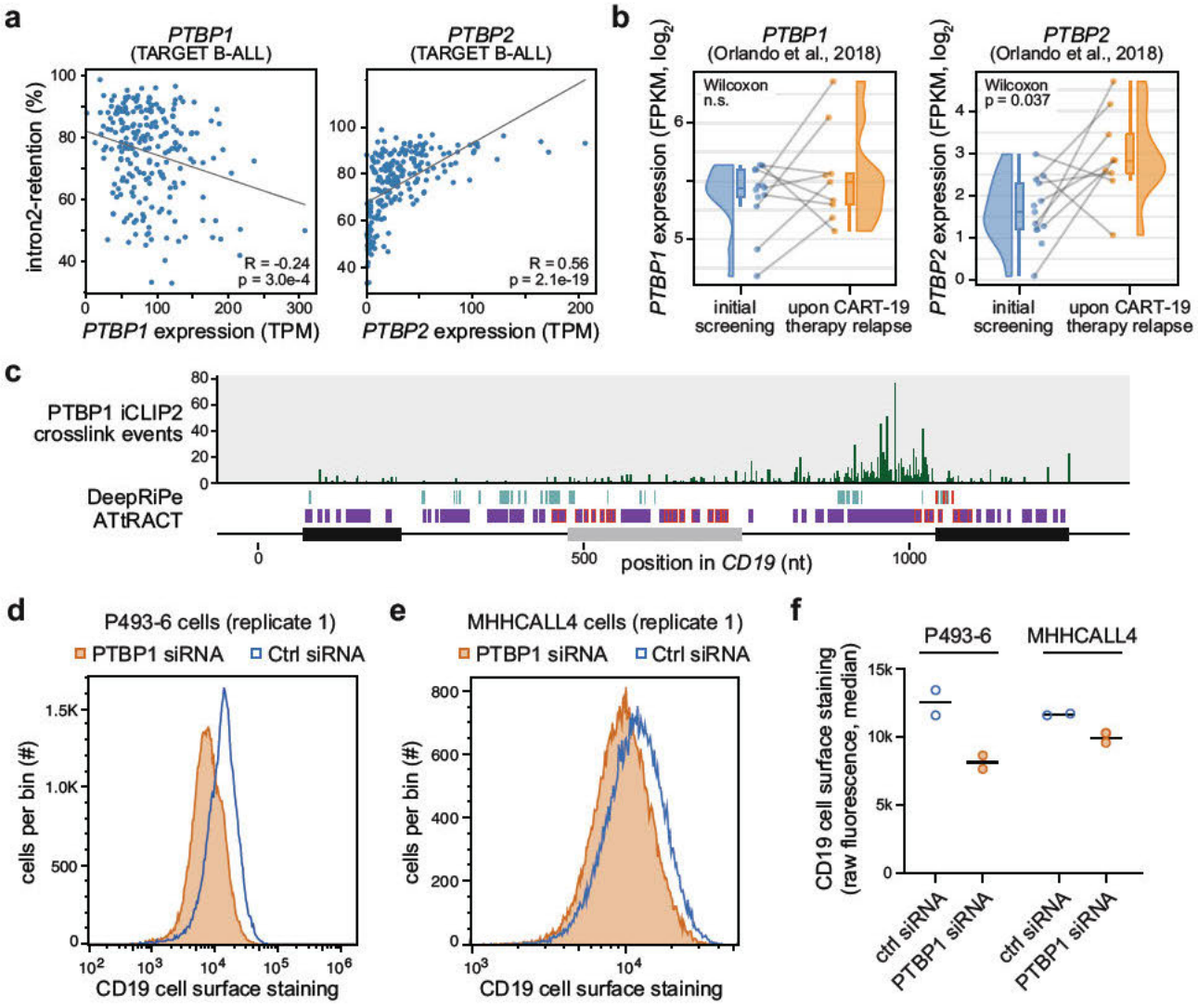


Fig. 6 PTBP1 is a regulator of *CD19* alternative splicing. a *PTBP1* and *PTBP2* mRNA levels correlate with *CD19* intron2-retention. Scatterplots comparing mRNA levels (TPM) to intron2-retention frequency for 220 B-ALL patient samples from TARGET B-ALL data. Pearson correlation coefficients and *P* values (two-sided) are given. **b** *PTBP2* mRNA levels are increased upon CART-19 therapy relapse. Box and violin plots showing *PTBP1* and *PTBP2* mRNA expression (fragments per kilobase of transcript per million mapped reads, FPKM) for patient samples ($n = 9$) from initial screening and upon CART-19 therapy relapse[5]. Grey lines connect matched samples of the same patients before therapy and after relapse. Boxes represent quartiles, centre lines denote 50th percentile, and whiskers extend to most extreme values within 1.5× IQR. One-sided paired Wilcoxon signed-rank test. **c** PTBP1 shows extensive binding to *CD19* intron 2. Bar diagram shows the number of PTBP1 iCLIP2 crosslink events from NALM-6 cells on each nucleotide in endogenous *CD19* exons 1–3. Predicted PTBP1 binding motifs (ATtRACT) and mutations predicted to alter PTBP1 binding (DeepRiPe) are shown below (see legend in Fig. 5d). Nucleotide positions are given relative to minigene sequence. **d**–**f** CD19 cell surface staining is reduced upon *PTBP1* knockdown in P493-6 (**d**) and MHHCALL4 (**e**) cells. Distributions of CD19 surface protein, as measured in 45–50 × 10^3 cells (per replicate) by CD19 antibody staining and flow cytometry, in cells transfected with *PTBP1* siRNA (orange) or non-targeting control siRNA (blue). **f** Dotplot shows mean and data points for measurements of cell surface CD19 in replicate 1 (**d**, **e**) and 2 (Supplementary Fig. 9f).

exon skipping, but tends to employ alternative splice sites and intron retention, both of which are clinically relevant in the case of CART-19 therapy resistance.

Our retrospective analyses of clinical B-ALL samples implicate unproductive *CD19* splice isoforms in the development of CART-19 therapy resistance. Using minigene assays, we directly show that *CD19* mutations that are observed in relapsed patients lead to exon 2 skipping, intron 2 retention or an additional isoform that uses an alternative 3′ splice site in exon 2. Furthermore, based on our mutational scan, we report ~200 additional point mutations that significantly affect these and other therapy-impacting isoforms. Thus, our results indicate that far more *CD19* mutations can create isoforms that would escape CART-19 recognition. In the future, targeted CRISPR/Cas9 replacement experiments using the endogenous *CD19* gene should be performed to validate that the predicted mutations cause physiological changes in splicing, loss of CD19 protein exposure on cell surface and CART-19 therapy resistance. Furthermore, the detection of such mutations in longitudinal samples may provide predictive biomarkers for therapy response in the future.

Likewise, alterations in the expression of *trans*-acting RBPs can induce aberrant *CD19* splicing, explaining the emergence of CD19-negative relapses in samples without mutations or with low-allelic-frequency mutations or without mutations in the *CD19* locus. Interestingly, we find that the differentiation status of B cells affects *CD19* splicing: in mature B cells, almost complete

exon 2 inclusion occurs, implying that all *CD19* transcripts give rise to functional CD19 protein. In contrast, intron 2 retention occurs in approximately half of the *CD19* transcripts in undifferentiated B-cell precursors (Fig. 1d). Likewise, retention of intron 2 is predominant in B-ALL patient samples from the TARGET B-ALL cohort, with 93% of patients exhibiting retention frequencies above 50% (Supplementary Fig. 1b). Hence, incomplete B-cell differentiation in B-ALL may induce a transcriptional and posttranscriptional programme, likely involving altered RBP expression, that reduces (but does not completely eliminate) the functional CD19 protein pool. This partial intron 2 retention may predispose the cancer cells to therapy resistance before they are actually subjected to CART-19 treatment, as observed in sorted B-cell populations from a B-ALL patients before and after CART-19 therapy relapse[17]. For the development of complete CART-19 resistance, some B-ALL patients thus likely host subclonal CD19-negative B-ALL cells which are further selected under the treatment[17]. The causes of complete CD19 loss in these subclonal cell populations are likely to be manifold, involving (epi)genetic changes such as hypermethylation of the *CD19* promoter[41], mutations in the *CD19* gene, splicing factor expression and combinations thereof.

Mutations in splicing factors such as SRSF2, SF3B1 and U2AF1 are common in myelodysplastic syndrome/acute myelogenous leukemia[42] and chronic lymphocytic leukemia[43], and are associated with aberrant splicing. In B-ALL, mutations in splicing factors are not common, but previous work suggests that splicing factor expression is deregulated[44]. In the context of *CD19*, we confirm that SRSF3 deregulation induces exon 2 skipping[9] and identify several other RBPs that promote the expression CD19 protein isoforms invisible to the immunotherapeutic agent, including PTBP1, PCBP2, SF3B4, HNRNPK, MBNL1 and HNRNPM. Several of the newly identified regulators have been found as deregulated in other cancer types and are discussed as potential targets for anti-cancer therapy[45–47]. In addition, upregulation of PTBP1 has been implicated in acquired resistance to the chemotherapeutic agent gemcitabine in pancreatic ductal carcinoma cells[48]. In the context of lymphocytes, PTBP1 is upregulated in B cells and required for early B-cell selection[49]. It was reported, however, that treatment of leukemic cells with the targeted therapy drug imatinib, which inactivates the BCR-ABL kinase encoded by the translocated Philadelphia (Ph) chromosome, lowers PTBP1 levels[50]. In the light of our finding that *PTBP1* knockdown increases *CD19* intron 2 retention and thereby reduces CD19 epitope presentation, previous treatments with imatinib may have negative impact on the subsequent response to the CART-19 therapy in a subset of Ph+ B-ALL patients. In addition, a recent study showed that the repeat RNA *PNCTR* sequesters substantial amounts of nuclear PTBP1 in various cancers[51]. Thus, in addition to regulation of PTBP1 expression, other factors such as availability may also influence PTBP1-mediated regulation in B-ALL cells under CART-19 therapy.

Currently, we cannot predict which patients with a CD19-positive B-ALL have a high risk of developing CD19-negative relapses. The pre-existence of isoforms skipping exon 2 or exons 5–6 has been previously discussed as a possible biomarker[16,17]. Moreover, in a comparison of B cells from a B-ALL patient, it was found that intron 2 retention had already occurred prior to CART-19 therapy (CD19-positive B cells) and had become predominant in the CD19-negative B cells after relapse[17]. Our results point to the need to extend the analysis to additional *CD19* isoforms and to incorporate the expression of splicing factors in screening approaches to identify patients at risk of relapse on CART-19 therapy. Notably, the same biomarkers might also be relevant for other malignancies arising from B-cell lineage, such as large B-cell lymphoma. Loss of CD19 following CART-19 therapy has been described as a mechanism for relapse[52], accounting for 60% of relapses in recent clinical studies[53]. Our data show that *CD19* splicing is highly complex, with already ~100 alternative isoforms concerning just exons 1–3. Of them, ~80% encode for a CD19 protein lacking a functional CART-19 epitope and are thus expected to contribute to therapy resistance. The specific detection of alternative splicing might serve as a reliable biomarker and may provide a novel approach to monitor disease progression as already suggested in other tumour entities[54]. To assess the role of the predicted cryptic splice isoforms in patients, we screened sequencing data from the TARGET B-ALL cohort and indeed recurrently found two junctions from the cryptic isoforms that we had observed in the mutagenesis data. Even though other cryptic junctions were absent and mutations associated with cryptic isoforms according to screen were also not found in the patient data, these may still emerge during CART-19 selection. Currently, there is a shortage of large-scale sequencing data of patient material before and after CART-19 therapy[55]. Future analysis of such data with a special focus on cryptic splice site usage will be important to identify mutations or splice isoforms that are predictive for CART-19 therapy success.

The contribution of aberrant splicing to CART-19 resistance may further be relevant for future combination therapies. Small-molecule splicing modulators are currently in clinical trials for myeloid neoplasms and splice-switching antisense oligonucleotides are in development for different targets (reviewed in[12]). Our mutagenesis dataset provides a strong basis for designing and systematically evaluating splice-switching oligonucleotides for the modulation of *CD19* splicing. The combined application of these splicing modulators with immunotherapy may represent a way to limit the generation of resistance to CART therapies.

Methods

Cell lines. NALM-6 cells were obtained from ATCC and cultured in RPMI medium (Life Technologies) with 10% fetal bovine serum (Life Technologies) and 1% L-glutamine (Life Technologies). HEK293T cells were obtained from DSMZ and grown with the same additives as for NALM-6. For validation experiments (Fig. 3e), HEK293 cells were obtained from DSMZ and were cultured in Gibco Dulbecco's Modified Eagle Medium (DMEM, Thermo Fisher Scientific) with L-Glutamine + 10% Gibco foetal bovine serum (FBS, Thermo Fisher Scientific). All cells were kept at 37 °C in a humidified incubator containing 5% CO_2. They were routinely tested for mycoplasma infection.

Cloning. The *CD19* minigene was amplified from human genomic DNA (Promega) with the primers 5′-catAAGCTTgaccaccgccttcctctctg-3′ and 5′-catGAATTCNNNNNNNNNNNNNNGGATCCttcccggcatctccccagtc-3′. pcDNA3.1 was used as the vector backbone for the *CD19* minigene plasmid. Both the backbone as well as the minigene amplicons were digested with the restriction enzymes *EcoR*I and *Hind*III (New England Biolabs). The backbone was extracted from a 1% agarose gel using QIAquick Gel Extraction Kit (Qiagen) and the minigene insert was cleaned up using QIAquick PCR Purification Kit (Qiagen). Ligation was conducted overnight at 16 °C with T4 DNA Ligase (New England Biolabs). All minigene mutations were introduced via Q5 Site-Directed Mutagenesis Kit (New England Biolabs). Position 748 (nucleotide 6 of intron 2) was exchanged from G to T to raise the baseline level of exon 2 inclusion in the WT *CD19* minigene to a similar level as in the endogenous *CD19* gene. The nine mutations from eight patients in Orlando et al.[5] listed in Supplementary Table 1 were inserted into the WT *CD19* minigene. For validation experiments (Fig. 3e), 19 individual point mutations with predicted effects on at least one isoform were inserted into a *CD19* minigene variant that additionally contained the mutation G742C to adjust the baseline of splice isoforms in HEK293 cells to the pattern seen in NALM-6 cells. All kits were used according to the manufacturers' recommendations.

Mutagenesis of minigene and library construction. For the random mutagenesis of the *CD19* minigene, GeneMorph II Random Mutagenesis Kit (Agilent) was used according to manufacturer's recommendations using 500 ng *CD19* minigene for 30 cycles at 56 °C with the amplification primers 5′-catAAGCTTgaccaccgccttcctctctg-3′ and 5′-catGAATTCNNNNNNNNNNNNNNGGATCCttcccggcatctccccagtc-3′. PCR products were purified using QIAquick Gel Extraction Kit (Qiagen), digested with *EcoR*I and *Hind*III (New England Biolabs) and then ligated into the backbone.

Transfection with minigenes. Cells were twice washed in Dulbecco's phosphate-buffered saline (DPBS, Gibco Thermo Fisher Scientific) and then collected in R buffer with a density of 2×10^7 cells/ml. For electroporation, we used 5 µg plasmid DNA (with a concentration of at least 1 µg/µl) to 2×10^6 cells in R buffer for a 100 µl NEON electroporation pipette tip (Thermo Fisher Scientific) at 1600 V for 30 ms and 1 pulse. Cells were harvested 24 h later. For validation experiments (Fig. 3e), 7×10^5 cells per well were seeded in six-well TC plates one day prior to transfection. 24 h later, cells were transfected with a mixture of 2 µg plasmid DNA and 20 µg linear Polyethylenimine MW 2500 (PEI 2500, Polysciences), and Gibco Opti-MEM (Thermo Fisher Scientific) was added to 100 µl total volume. This mixture was added dropwise to 1,9 ml fresh DMEM covering the cells, followed by incubation for 24 h.

Quantification of splicing isoforms using semiquantitative RT-PCR. Semi-quantitative RT-PCR was used to quantify ratios of *CD19* mRNA isoform variants. To this end, reverse transcription was performed on 500 ng RNA with RevertAid Reverse Transcriptase (Thermo Fisher Scientific) according to the manufacturer's recommendations. Subsequently, 1 µl of the cDNA was used as template for the RT-PCR reaction with OneTaq DNA Polymerase (New England Biolabs). PCRs were run at the following conditions: 94 °C for 30 s, 28 cycles (minigene) or 34 cycles (endogenous *CD19*) of [94 °C for 20 s, 55 °C for 30 s, 68 °C for 30 s] and final extension at 68 °C for 5 min. The primers 5′-ACCTCCTCGCCTCCTCTTCTTC-3′ and 5′-GCAACTAGAAGGCACAGTCG-3′ were used for the *CD19* minigene, and 5′-ACCTCCTCGCCTCCTCTTCTTC-3′ and 5′-CCGAAACATTCCACCGGAA CAGC-3′ for the endogenous *CD19* gene. A TapeStation 2200 capillary gel electrophoresis instrument (Agilent) was used for quantification of the PCR products on D1000 tapes.

For the semiquantitative RT-PCR experiments in HEK293 cells, cells were harvested 24 h after transfection and pelleted. RNA was isolated using the QiaGen RNeasy kit following the manufacturer's protocol with the exception of adding only 100 µl of cell lysate onto the gDNA removal columns to ensure proper removal of genomic and/or plasmid DNA. 1–2 µg RNA per sample were used to generate cDNA with the ThermoScientific RevertAid cDNA kit. We changed the second temperature step of the manufacturer's synthesis protocol from 4 °C to 25 °C (5 min) to further reduce RNA dimerisation or formation of secondary structures. Subsequently, 1 µl of cDNA was used as template for the RT-PCR reaction with OneTaq DNA Polymerase (New England Biolabs). PCRs were run at the following conditions: 94 °C for 30 s, 28 cycles of [94 °C for 20 s, 52 °C for 30 s, 68 °C for 30 s] (minigene) or 34 cycles of [94 °C for 20 s, 55 °C for 30 s, 68 °C for 30 s] (endogenous CD19) and final extension at 68 °C for 5 min. The primers 5′-ACCTCCTCGCCTCCTCTTCTTC-3′ and 5′-GCAACTAGAAGGCACAGTCG-3′ were used for the CD19 minigene, and 5′-ACCTCCTCGCCTCCTCTTCTTC-3′ and 5′-CCGAAACATTCCACCGGAACAGC-3′ for the endogenous CD19 gene. The PCR products were quantified using the TapeStation 4200 system and the High Sensitivity D1000 reagents and tapes (Agilent) according to the manufacturer's protocol.

Significance of differences in isoform abundance for comparing WT minigenes vs. mutated variants (Fig. 1i) or RBP knockdown vs. control (Fig. 5f) was calculated by a Student's *t* test separately for each isoform, reporting the smallest *P* value for each comparison. A one-way ANOVA was used to test whether isoform abundances are different in any of the three conditions shown in Fig. 1g.

Generation of stable and inducible shRNA knockdown cell lines

Production and preparation of lentivirus. Oligonucleotides with shRNA inserts against eleven RBPs (Supplementary Table 2) were ordered as Ultramer DNA Oligos from Integrated DNA Technologies (Leuven, Belgium). All sequences were based on[56]. Oligonucleotides containing shRNA inserts were PCR-amplified with primers 5′-TCTCGAATTCTAGCCCCTTGAAGTCCGAGGCAGTAGGC-3′ and 5′-TGAACTCGAGAAGGTATATTGCTGTTGACAGTGAGCG-3′ and purified with QIAquick PCR Purification Kit (Qiagen). shRNA inserts and miR-E18_LT3GEPIR_Ren714 backbone (inducible via Tet-On system) were cut with *Eco*RI and *Xho*I (New England Biolabs). Backbone was purified from agarose gel with QIAquick Gel Extraction Kit (Qiagen). The fragments were then ligated with T4 DNA Ligase (New England Biolabs) at 16 °C overnight.

Constructs were transduced into NALM-6 via HEK293T-produced lentiviruses. To this end, 10 cm dishes of HEK293T were transfected using 30 µl Lipofectamine 2000 (Thermo Fisher Scientific) with three plasmids: 4 µg shRNA-producing constructs + 2 µg psPAX2 (lentiviral packaging) + 1 µg pMD2.G (lentiviral envelope) at 72 h prior to transduction. On the first day after transfection, the medium was changed. Work with cells used for lentiviral production was conducted in the S2 laboratory.

Transduction of NALM-6 cells. Lentiviral production was confirmed with Lenti-X GoStix (Takara) and lentiviruses were concentrated with Lenti-X Concentrator (Takara) according to the manufacturer's recommendations. For transduction, 1×10^6 NALM-6 cells in 500 µl of medium were added to the concentrated virus. 5 µg/ml polybrene (Sigma-Aldrich) was added. The cells were centrifuged at 800 g and 32 °C for 30 min. Cells were then transferred into 6-well plates and cultivated in normal growth medium without antibiotics. Selection was started after 48 h with 0.5 µg/ml puromycin (Thermo Fisher Scientific). Antibiotic medium was exchanged every 2–3 days. As soon as cells were not dying under selection anymore and the population was stable, induction experiments were started. After transduction, cells remained in the S2 laboratory for at least 6 weeks. Then, Lenti-X GoStix was used to check for any remaining lentivirus.

Induction of stable shRNA-expressing NALM-6 cells. Controlled by the Tet-responsive *TRE3G* promoter, the expression of shRNA was induced by addition of doxycycline (Thermo Fisher Scientific). To this end, 2×10^6 NALM-6 cells were seeded into a six-well plate in 2 ml medium containing 0.5 µg/ml puromycin and induced with 0.5 µg/ml doxycycline, diluted in RPMI 1640 medium (Thermo Fisher Scientific). Induction was conducted at 37 °C and 5% CO_2 and cells were harvested after 48 h. During induction, the shRNA expression system is coupled to the production of eGFP, which was examined by fluorescence microscopy before harvesting.

Quantitative real-time PCR (qPCR). RNA was extracted from the induced harvested cells using the RNeasy Plus Mini Kit (Qiagen). This RNA was used for qPCR to validate the RBP knockdown as well as for semiquantitative RT-PCR experiments to check the splicing pattern of endogenous *CD19*. The qPCR was conducted using the Luminaris HiGreen qPCR Master Mix, low ROX (Thermo Fisher Scientific) according to the manufacturer's recommendations. Oligonucleotide sequences of all qPCR primers are given in Supplementary Table 3.

Targeted DNA sequencing. DNA-seq of the minigene library was performed on the PacBio SMRT sequencing platform at MPI-CBG Dresden. For this purpose, the minigene plasmid library was digested with *Eco*RI and *Hind*III (New England Biolabs) and run on an agarose gel. The desired band at the size of 1301 nt was cut out and purified using QIAquick Gel Extraction Kit (Qiagen). For the run on the PacBio SMRT cell, standard library preparation was performed.

Targeted RNA sequencing. NALM-6 cells were electroporated with the mutated minigene library (see above). 24 h later cells were harvested and RNA was isolated via the RNeasy Mini Kit (Qiagen). 20 µg isolated RNA was poly-A-selected using Dynabeads Oligo $(dT)_{25}$ beads (Invitrogen) according to the manufacturer's recommendations. Reverse transcription was performed on 500 ng poly-A-selected RNA with RevertAid Reverse Transcriptase (Thermo Fisher Scientific) according to the manufacturer's recommendations. To prevent chimeric amplicons, the RNA-seq libraries were amplified via emulsion PCR[57] using the Phusion DNA Polymerase (New England Biolabs). The following primers containing Illumina adaptors were used in the PCR: 5′- CAAGCAGAAGACGGCATACGAGATCGGTCTC GGCATTCCTGCTGAACCGCTCTTCCGATCTNNNNNNNNNNGGAACCTCT AGTGGTGAAGG-3′ (fwd) 5′-AATGATACGGCGACCACCGAGATCTACACT CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNCCGCCAGTGTG ATGGATATC-3′ (rev) under following conditions: 98 °C for 30 s, 25 cycles of [98 °C for 10 s, 63 °C for 20 s, 72 °C for 1 min] and final extension at 72 °C for 5 min. Amplicons were purified using Agencourt AMPure XP beads (Backman Coulter). Purified products were analysed on the TapeStation 2200 capillary gel electrophoresis instrument (Agilent) and quantified using the Qubit assay (Thermo Fisher Scientific). RNA-seq was carried out on the Illumina MiSeq platform using paired-end reads of 350 nt + 250 nt length and a 10% PhiX spike-in to increase sequence complexity.

PTBP1 iCLIP2 experiments. We used the iCLIP2 approach for transcriptome-wide mapping of PTBP1 binding to RNA in NALM-6 cells. iCLIP2 was performed according to our previously published protocol[35]. Briefly, the iCLIP2 libraries were made from NALM-6 cells grown in RPMI as described above (2×10^6 cells per replicate). To induce protein-RNA crosslinks, the cells were irradiated with 150 mJ/cm^2 UV light at 254 nm. Next, PTBP1 protein-RNA complexes were immunoprecipitated using 2 µg of anti-PTBP1 antibody (Santa Cruz, sc-56701). RNase digestion was performed by adding 10 µl of 1/300 or 1/500 diluted RNase I (Ambion) to the sample. RNA purification, reverse transcription and library preparation were done as described in[35].

PTBP1 siRNA electroporation in MHHCALL4 and P493-6 cells. The cell lines MHHCALL4 and P493-6 were electroporated with a specific siRNA targeting PTBP1 (TAGCAAGATGATACAATGGTA[dT][dT]; Sigma, sR90) or a Scramble control (D-001810-10-50, Dharmacon) using the Neon Transfection System (Thermo Fisher Scientific). In short, 5×10^5 cells were resuspended in 10 µl of 5 µM siRNA in buffer R and electroporated using the Neon Transfection System 10 µL Kit (MPK1096, Thermo Fisher Scientific) with the following settings: 1700 V, 20 ms, 1 pulse. After electroporation, the cells were cultured in the recommended media for 48 h and collected for CD19 cell surface staining, quantitative real-time PCR and Western blot.

CD19 cell surface staining. In all, 1×10^5 cells were resuspended in 50 µl of PBS, 20% FBS, 1 mM EDTA and 2.5 µl of Human TruStain FcX blocking (422302, BioLegend) and incubated for 20 min. After blocking, 2.5 µl of APC anti-human CD19 antibody (1:20, 982406, BioLegend) was added to the cells and incubated for 30 min. Cells were washed twice with PBS, 20% FBS, 1 mM EDTA and the CD19 staining was measured using the BD Accuri C6 Plus Flow Cytometer

instrument (BD Biosciences). Flow cytometry data was analysed with FlowJo (version 10.7.2) software.

Western blot. Cell pellets were resuspended in RIPA buffer with Protease/Phosphatase Inhibitor (1861282, ThermoScientific) and 30 µg of protein were loaded in a 10% pre-cast gel (456-1035, BioRad). Antibodies against CD19 (1:1000, #3574, Cell Signalling), PTBP1 (1:500, sc-56701, Santa Cruz Biotechnology) and β-Actin (1:5000, 8H10D10, Cell Signalling) were used for total protein expression detection. Images were acquired with GBox instrument (Syngene).

Quantitative real-time PCR. RNA was extracted using the Maxwell RSC Instrument (Promega) and the Maxwell® RSC simplyRNA Cells Kit (AS1390, Promega). 0.5 µg or RNA were reverse-transcribed using the SuperScript™ IV Reverse Transcriptase kit (18090010, Invitrogen) following the manufacturer's protocol. RNA expression was measured using SYBR Green Master MIX (Thermo Fisher Scientific). Quantitative real-time PCR was performed in a QuantStudio™ 7 Pro Real-Time PCR System (Thermo Fisher Scientific) with specific primers (Supplementary Table 3) spanning the exon-exon junctions between exons 2 and 3 (E2E3 1&2), 3 and 4 (E3E4), and 10 and 12 (E10E12) as well as the exon-intron junctions from exon 2 to intron 2 (e2i2), from intron 2 to exon 3 (i2e3).

Re-analysis of RNA-seq data from Orlando et al. We re-analysed RNA-seq data of B-ALL patients at screening and after CART-19 therapy relapse from Orlando et al.[5] to quantify intron 2 retention in *CD19*. Since raw data were not available, we obtained BAM files for the different patients deposited in the Short Read Archive (SRA) under the accession SRP141691. For 10 patients, matched data were available at screening and relapse. We excluded one patient (patient #17) after visual inspection indicating that the submitted data in fact corresponded to DNA-seq rather than RNA-seq data. The data contained the aligned reads mapped to several genes from the immune system including *CD19*. Using custom scripts, we extracted the sequence of the reads, reformatted them and generated fastq files. We then mapped the fastq files to our minigene sequence using STAR[58] (v2.6.1). We used the re-mapped reads to quantify the levels of intron 2 retention in the different samples using the R/Bioconductor package ASpli[59] (version 1.12.1).

For the expression analysis of B-ALL patients at screening and after CART-19 therapy relapse, we used the gene read counts provided in Supplementary Data Table 1 of Orlando et al.[5]. Gene lengths were taken from BiomaRt (version 2.4.21) for the human genome version GRCh37 accessed through the R/Bioconductor package OrgDb (version 3.10.0). Normalisation and RPKM calculations were performed using the R/Bioconductor package edgeR[60] (version 3.28.1).

DNA-seq barcode demultiplexing. We obtained the circular consensus sequences (CCS), stored as fastq files. Two rounds of sequencing yielded a total of 337,215 CCS. We kept only reads with a length of 150–1150 nt. We adapted the demultiplexing procedure described in[19]. In this case, we searched for the 15-nt barcode in the last 50 nt of the read. If the barcode was not found, we searched in the last 50 nt of the reverse complementary strand. We only allowed the recovery of barcodes ranging from 14 to 16 nt, which would account for barcodes containing one nucleotide inserted or deleted. Before proceeding with the variant calling, we determined a cutoff to decide the minimal number of CCS to call variants on. Here, we kept only barcodes supported by at least 4 CCS. In total, we recovered 68.5% of all the demultiplexed barcodes which corresponded to 10,558 different minigenes, closely resembling the ~10,000 minigene clones that were used to generate the library.

DNA-seq mapping and variant calling. We use BLASR[61] with the standard parameters to map the demultiplexed minigene sequences to the minigene reference. We performed variant calling in the aligned BAM files using the GATK[62] HaplotypeCaller (version 4.0.10) with the parameters *--kmer-size 10 --kmer-size 15 --kmer-size 25 --allow-non-unique-kmers-in-ref*. We used different *k*-mer sizes to improve the detection of problematic regions. Mixed barcodes, i.e., barcodes containing two classes of mutations, were removed based on the "penetrance score", reported as allele frequency (AF) in the GATK vcf output files, such that barcodes with more than 25% variants of low penetrance (AF < 0.8) were discarded. Using this strategy, we were able to recover 100,135 mutations of high quality coming from 10,295 distinct minigenes plus an additional 194 unmutated WT minigenes with distinct barcodes. 57.4% of the mutations appeared in at least ten different minigenes.

RNA-seq barcode demultiplexing. RNA-seq libraries were sequenced on Illumina MiSeq as 350 nt + 250 nt paired-end reads, yielding approximately 23 million reads. We controlled their quality using FastQC (version 0.11.5, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and removed bad quality ends of reads using Trimmomatic[63] (version 0.36, parameters: SLIDINGWINDOW:6:10 MINLEN:0). After trimming, we filtered for read pairs with a minimal length of 305 nt (read1) and 157 nt (read2) and, as done in Braun et al.[19], we used matchLRPatterns() and trimLRPatterns() from the R/Bioconductor package Biostrings to extract the 15-nt barcode in read1 between the two flanking restriction sites (Lpattern = TGCAGAATTC, Rpattern = GGATCC) allowing one mismatch. All read pairs with barcode length between 14 and 16 nt were kept for further processing. Barcode sequences were added to the read names in the fastq file and 5′ ends of reads were trimming (read1: everything until the second anchor sequence GGATCC, read2: the first 12 nt). After identifying and trimming the barcode and other regions, we used Cutadapt[64] (version 1.6, parameters: --adaptor=TAGAGGTTCC --overlap=3 --error-rate=0.1 --no-indels --minimum-length=244 --pair-filter=both) to remove remaining primer sequences from read1. Lastly, the barcode information attached to the read names was used to demultiplex all read pairs into individual fastq files for each minigene.

Isoform quantification from RNA-seq data. Only barcodes/minigenes also detected in the DNA-seq library were kept for further analysis. All minigenes with insertions or deletions of 10 or more base pairs were removed from further analysis. For better mapping results, we shortened read1 to at most 260 nt. Read pairs of each minigene were mapped to the respective minigene (including all mutations with penetrance ≥0.8, but excluding insertions and deletions) using STAR[58] (version 2.6.1b). An annotation of three isoforms (exon 2 inclusion and skipping, as well as the artefact PCR product Δex2part which lacks an internal fragment of exon 2 due to a reverse transcription artefact[65]) was provided to STAR during mapping and an --sjdbOverhang of 259 was set. When running STAR, all SAM attributes were written, up to ten mismatches were allowed, soft-clipping was prohibited on both ends of the reads and only uniquely mapping reads were kept for further analysis. BAM files were sorted and indexed using SAMtools[66] (version 1.5).

Properly and consistently mapped pairs were used for isoform reconstruction using a custom Perl script. Read pairs were considered properly mapped if they mapped with the right orientation on opposite strands. Read pairs mapped consistently if they either did not overlap or in case of an overlap, agreed in their detected splice junctions. Besides, only read pairs for which both mates exceeded the constitutive exon boundaries by at least 10 nt were used for isoform reconstruction. All other pairs were removed since they did not provide any isoform information. Only minigenes covered by at least 100 read pairs usable for isoform reconstruction were kept for further analysis. For each read pair, the CIGAR strings of the two mates were used to reconstruct their splicing isoform. Regarding the artefact product Δex2part, we combined the eight possible mappings of the missing internal fragment of exon 2 which are possible due to the associated 8-nt repeat sequence[65]. Only isoforms, which were supported by ≥ 1% of the read pairs and at least two read pairs in at least one minigene, were kept for further analysis.

The analysis described above was done separately for two replicates. All isoforms occurring with a frequency of at least 5% in two or more minigene variants in either of the two replicates were kept as individual isoforms. All other detected isoforms were summarised into a category "discarded". Isoforms with Δex2part, i.e., excluding the internal intron in exon 2, were combined with their "real" counterparts without Δex2part by merging isoforms that only differed in the exclusion of the internal fragment of exon 2. In total, this leads to a set of 101 individual isoforms.

Estimation of single-mutation effects and splicing-affecting mutations. Since the majority of the minigenes in the dataset exhibit more than one mutation, with a mean of 9.6 mutations per minigene, the splicing-affecting mutations cannot be read out directly from the data. We used multinomial logistic regression to infer the effects of single mutations from combined measurements. The regression is based on hypothetical minigenes containing only one mutation, and on the assumption that mutation effects (log fold-changes compared to WT) add up into combined ones at the levels splice isoform ratios[19].

For regression, we focused on the five major isoforms that are already present in the WT minigene (see main text). Therefore, minigenes exhibiting more than 5% cryptic isoforms were removed from the dataset, and for the remaining minigenes the cryptic isoforms were merged into a lumped splicing category which we termed "other". Thus, six categorical splicing outputs (inclusion, skipping, intron2-retention, alt-exon2, alt-exon3, other) were considered in the regression model, and the probability of each these outputs to be observed was assumed to equal the measured isoform frequencies. The regression was formulated as a softmax regression problem using the LogisticRegression command from the Python package scikit-learn[67].

Given the large number of mutations per minigene in the dataset, the regression was prone to overfitting (i.e., mutations with weak effects on splicing were assigned non-zero coefficients to fit random fluctuations in the data; not shown). To avoid this problem, we employed L1 penalisation. The strength of the penalty was optimised by tenfold cross-validation, and the resulting inverse regularisation strength was C = 10 for both replicates.

The goodness of the model in describing the measured combined mutation effects (minigenes) was tested by assessing the correlation between model and data in training and test datasets (Supplementary Fig. 4a). Tenfold cross-validation was performed by once randomly splitting the dataset into ten parts. In ten distinct model evaluations, nine of the data sections were simultaneously for model training, whereas the remaining section served as test data. Therefore, each data point is only once part of the test data, and the mean Pearson correlation coefficient between model prediction and test data was used to assess the model performance. Cross-validation at the final penalisation strength (with the highest correlation between model and test data) showed that the method performs very well in estimating the minigene isoform frequencies of the test dataset

(Supplementary Fig. 4b). Since we saw little variability in the prediction power for these ten validation runs, we report the average correlation coefficient in Fig. 3b: The Pearson correlation coefficients between softmax predictions of combined mutation effects and measurements lie for the single isoforms between 0.68-0.95 for the first replicate and between 0.71-0.93 for the second replicate.

The accuracy of the model-predicted single-mutation effects in the softmax regression was assessed by leaving out 56 directly measured single-mutation minigenes (i.e., minigenes bearing only one mutation) from the training data. Since most of these 56 mutations are not splicing-affecting, we focused our analysis on the seven mutations that change the inclusion isoform level beyond two standard deviations of the WT minigene distribution: For each of the seven mutations, we performed multiple softmax fits in which the training data: (i) contained all minigenes not harbouring the mutation of interest, (ii) excluded its single-mutation minigenes, and (iii) comprised varying numbers of combined mutation minigenes containing the mutation. For each mutation occurrence between 1 and 10, we used up to 7 different, randomly chosen combinations of multiple mutated minigenes including the mutation of interest. For each of these models, we generated predictions for the single-mutation effect. The prediction accuracy was assessed by calculating the difference between model and direct single-mutation measurements for a certain mutation occurrence. The standard deviation of the difference between model and data was used as a measure for the model error. We find that a mutation occurrence of 3 leads to an error level equal to two WT standard deviations (calculated based on inclusion levels of all WT minigenes in the first replicate). For higher mutation occurrences, the prediction accuracy does not improve further (Supplementary Fig. 4c).

The final modelling step was to identify splicing-affecting mutations. For this purpose, we adopted an approach analogous to empirical *P* values, i.e., we compared predicted single-mutation effects to empirical isoform frequency distributions in the WT. Isoform frequencies were measured for 195 and 194 WT minigenes in the two replicates. For each isoform and replicate, we chose the 2.5% and 97.5% quantiles of the respective empirical WT frequency distribution as cutoffs (corresponding to a two-sided 5% cutoff). A mutation was considered to have an effect on a splice isoform if, for both replicates, the frequencies predicted by the model were beyond the respective cutoffs and if the effects were in the same direction.

Splice site characterisation. Splice site usage for a given position represents the frequency of the isoforms using a given splice site in a particular minigene divided by the sum of all isoform frequencies for the same minigene. For Fig. 4a, we used the maximum usage of a particular splice site across all minigenes. The strength of putative splice sites along the minigene was calculated using MaxEnt scores[26] in sliding windows of 9 nt or 23-nt to evaluate the corresponding sequences as potential 5′ or 3′ splice sites, respectively. The procedure was repeated for all individual point mutations to assess their potential to create cryptic splice sites. For the calculations, we used the Python implementation of MaxEnt (maxentpy, version 0.0.1, https://github.com/kepbod/maxentpy). We filtered the output by keeping only windows that contained a GU or AG dinucleotide in the positions 4–5 (5′ splice site) or 19-20 (3′ splice site), respectively.

We compared the effects of single-point mutations in our library to predictions by the state-of-the-art deep learning algorithm SpliceAI[25]. We ran SpliceAI (version 1.3.1) with the default parameters plus masking (-M1), using GENCODE[68] (v31) annotation for the human genome version hg38 as a reference. Given that SpliceAI results are reported in terms of a probability of gain or loss of a particular splice site, we assigned the gained splice sites in our cryptic isoforms by comparison to the canonical exon 2 inclusion isoform, such that if a new splice site appears in the cryptic isoform, it is considered as "gained" with respect to the "lost" WT splice site. All splice sites in a cryptic isoform were given the same prevalence score, i.e., the prevalence score of the mutation-isoform pair. To compare the SpliceAI scores for a given splice site gain with our prevalence score (Fig. 4f), we considered the mutations that (i) share the same gain-loss pair of positions in both assays, and (ii) are predicted by SpliceAI to gain of a new splice site (i.e., a cryptic site where score_gain > score_loss) upon a given mutation.

RBP binding site predictions. For the prediction of RBP binding motifs, we used the web versions of the oRNAment[28] (http://rnabiology.ircm.qc.ca/oRNAment) and ATtRACT[27] (https://attract.cnic.es/) databases to query the minigene sequence for presence of RBP motifs (Supplementary Fig. 7a). From the obtained predictions, we collapsed overlapping binding sites from the same tool and RBP.

We used DeepRiPe[30] to predict the potential impact of single-point mutations on RBP binding. To this end, we downloaded the trained models for PAR-CLIP and ENCODE eCLIP data on 159 RBPs available in the GitHub repository (https://github.com/ohlerlab/DeepRiPe). We scored each mutation (annotated with regards to the hg38 reference genome) across the individual RBP models and reported every mutation for which the model score changed by at least 0.1 in either direction compared to the WT sequence (Supplementary Data 6, worksheet "DeepRiPe mutations"). Positive and negative delta scores refer to a predicted increase or reduction in RBP binding, respectively. The scoring functions are based on the iPython notebooks provided by DeepRiPe: https://colab.research.google.com/drive/18yeqRE7KmOjfbUaLAfJ6rMBjAulYo-Uc?usp=sharing

For the definition of significant RBP binding sites, we used the following strategy. For binding sites predicted by oRNAment and ATtRACT, we first checked their overlap separately for each isoform. If a binding site overlapped in at least one position with a splicing-affecting mutation with respect to this particular isoform, we defined this binding site as an isoform-specific significant binding site. All binding sites that were significant for at least one isoform were collapsed into the complete list of significant binding sites, yielding a total of 315 significant binding sites for 74 RBPs. In the case of DeepRiPe, a mutation with a delta score >0.25 for a given RBP model was required to overlap with a splicing-affecting mutation for a particular isoform (our screen) to be considered an isoform-specific significant RBP-changing mutation. In a similar manner, all isoform-specific mutations for any isoform were collapsed into a complete list of significant RBP-changing mutations, yielding a total of 222 significant mutations that affected the binding of 58 RBPs.

iCLIP data processing. iCLIP libraries were sequenced on an Illumina NextSeq 500 sequencing machine as 92 nt single-end reads including a 6 nt sample barcode as well as 5 + 4 nt unique molecular identifiers (UMIs). Basic quality controls were done with FastQC (version 0.11.8) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and reads were filtered based on sequencing qualities (Phred score) in the barcode region using the FASTX-Toolkit (version 0.0.14) (http://hannonlab.cshl.edu/fastx toolkit/) and seqtk (version 1.3) (https://github.com/lh3/seqtk/). Reads were demultiplexed based on the experimental barcode, which is found on positions 6 to 11 of the reads, using Flexbar[69] (version 3.4.0). Afterwards, barcode regions and adaptor sequences were trimmed from read ends using Flexbar. Here, a minimal overlap of 1 nt of read and adapter was required, UMIs were added to the read names and reads shorter than 15-nt were removed from further analysis. Downstream analysis was done as described in Chapters 3.4 and 4.1 of Busch et al.[70]. Genome assembly and annotation of GENCODE[68] v31 were used during mapping.

Patient data analysis. RNA-seq data of 222 B-ALL patients from the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) programme (https://ocg.cancer.gov/programs/target) were processed from fastq files. Sequencing adaptors were trimmed with TrimGalore[71] (version 0.6.6), aligned to the hg38 human genome assembly with STAR[58] (version 2.5.2a), and sorted and indexed with SAMtools[66] (version 1.11). Splice junctions were quantified individually for each sample using MAJIQ[72] (version 2.2) and ENSEMBL reference transcriptome GRCh38.94[73]. Only splice junctions with a usage level (percent selected index, PSI) of at least 5% in any given TARGET B-ALL samples were quantified. The local splicing variation (LSV) harbouring alternative splicing in the region of the *CD19* minigene (Supplementary Fig. 1a) was quantified in 220 out of 222 B-ALL patients. For comparison, we used RNA-seq data of immature and mature B cells from healthy donors from[44,74].

Annotated variant call format (VCF) files were downloaded for the TARGET B-ALL patient cohort from the NCI Genomic Data Commons (GDC) Data Portal (accessed 11/18/2021). These files are available under controlled access (see Acknowledgements). In brief, the VCF files had been generated from patient whole-exome DNA-seq data using the GDC DNA-seq Analysis Pipeline (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/) which includes genomic alignment with BWA, data clean-up with Picard tools and GATK, and calling of somatic variants from matched samples of tumor and adjacent normal tissue for each patient with Mutect2. The raw VCF files were further annotated using the Variant Effect Predictor (VEP) tool to infer the location of each mutation, its consequence (frameshift/silent mutation) and the affected gene(s) as well its overlap with known variants in databases such as ClinVar and dbSNP. Using custom scripts, a total of 468 VCF files of TARGET B-ALL patients were parsed for *CD19* mutations. This identified 39 patients with somatic mutations within the *CD19* gene, including 11 mutations in the *CD19* minigene region (Supplementary Data 4).

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

Data availability

The sequencing data generated in this study have been deposited in the Gene Expression Omnibus (GEO) database under accession code GSE182894. The collection consists of the PacBio DNA-seq libraries (GSE182892) [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc GSE182891], the Illumina RNA-seq libraries (GSE182892) and the PTBP1 iCLIP2 libraries in NALM-6 cells (GSE182893). The results published here are in whole or part based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments (https://ocg.cancer.gov/programs/target) initiative, phs000218. The data used for this analysis are available at https://portal.gdc.cancer.gov/projects. The remaining data are available within the Article, Supplementary Information or Source Data files. Source data are provided in this paper.

Code availability

The computational code for the analyses and figure generation is available in Zenodo [https://doi.org/10.5281/zenodo.6614454]/Github [https://github.com/mcortes-lopez/CD19_splicing_mutagenesis] under an open-source MIT license.

Received: 7 October 2021; Accepted: 5 July 2022;
Published online: 22 September 2022

References

1. Maude, S. L. et al. Chimeric antigen receptor T cells for sustained remissions in leukemia. *N. Engl. J. Med.* **371**, 1507–1517 (2014).
2. Davila, M. L. & Brentjens, R. J. CD19-Targeted CAR T cells as novel cancer immunotherapy for relapsed or refractory B-cell acute lymphoblastic leukemia. *Clin. Adv. Hematol. Oncol.* **14**, 802–808 (2016).
3. Wudhikarn, K. et al. Interventions and outcomes of adult patients with B-ALL progressing after CD19 chimeric antigen receptor T-cell therapy. *Blood* **138**, 531–543 (2021).
4. Roberts, K. G. Genetics and prognosis of ALL in children vs adults. *Hematol. Am. Soc. Hematol. Educ. Program* **2018**, 137–145 (2018).
5. Orlando, E. J. et al. Genetic mechanisms of target antigen loss in CAR19 therapy of acute lymphoblastic leukemia. *Nat. Med.* **24**, 1504–1506 (2018).
6. Park, J. H. et al. Long-term follow-up of CD19 CAR therapy in acute lymphoblastic leukemia. *N. Engl. J. Med.* **378**, 449–459 (2018).
7. Gardner, R. A. et al. Intent-to-treat leukemia remission by CD19 CAR T cells of defined formulation and dose in children and young adults. *Blood* **129**, 3322–3331 (2017).
8. Maude, S. L. et al. Tisagenlecleucel in Children and Young Adults with B-Cell Lymphoblastic Leukemia. *N. Engl. J. Med.* **378**, 439–448 (2018).
9. Sotillo, E. et al. Convergence of acquired mutations and alternative splicing of CD19 enables resistance to CART-19 immunotherapy. *Cancer Discov.* **5**, 1282–1295 (2015).
10. Shah, N. N. & Fry, T. J. Mechanisms of resistance to CAR T cell therapy. *Nat. Rev. Clin. Oncol.* **16**, 372–385 (2019).
11. Asnani, M. et al. Retention of CD19 intron 2 contributes to CART-19 resistance in leukemias with subclonal frameshift mutations in CD19. *Leukemia* **34**, 1202–CD1207 (2020).
12. Bonnal, S. C., López-Oreja, I. & Valcárcel, J. Roles and mechanisms of alternative splicing in cancer - implications for care. *Nat. Rev. Clin. Oncol.* **17**, 457–474 (2020).
13. Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16**, 413–430 (2016).
14. El Marabti, E. & Abdel-Wahab, O. Therapeutic modulation of RNA splicing in malignant and non-malignant disease. *Trends Mol. Med.* **27**, 643–659 (2021).
15. Bagashev, A. et al. CD19 alterations emerging after CD19-directed immunotherapy cause retention of the misfolded protein in the endoplasmic reticulum. *Mol. Cell Biol.* **38**, e00383–18 (2018).
16. Fischer, J. et al. CD19 isoforms enabling resistance to CART-19 immunotherapy are expressed in B-ALL patients at initial diagnosis. *J. Immunother.* **40**, 187–195 (2017).
17. Rabilloud, T. et al. Single-cell profiling identifies pre-existing CD19-negative subclones in a B-ALL patient with CD19-negative relapse after CAR-T therapy. *Nat. Commun.* **12**, 865 (2021).
18. Zhao, Y. et al. Tumor-intrinsic and -extrinsic determinants of response to blinatumomab in adults with B-ALL. *Blood* **137**, 471–484 (2021).
19. Braun, S. et al. Decoding a cancer-relevant splicing decision in the *RON* proto-oncogene using high-throughput mutagenesis. *Nat. Commun.* **9**, 3315 (2018).
20. Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J. & Lehner, B. Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell* **176**, 549–563.e523 (2019).
21. Baeza-Centurion P., Miñana B., Valcárcel J. & Lehner B. Mutations primarily alter the inclusion of alternatively spliced exons. *Elife* **9**, e59959 (2020).
22. Ke, S. et al. Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res.* **28**, 11–24 (2018).
23. Glidden, D. T., Buerer, J. L., Saueressig, C. F. & Fairbrother, W. G. Hotspot exons are common targets of splicing perturbations. *Nat. Commun.* **12**, 2756 (2021).
24. Enculescu, M. et al. Exon definition facilitates reliable control of alternative splicing in the *RON* proto-oncogene. *Biophys. J.* **118**, 2027–2041 (2020).
25. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e524 (2019).
26. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
27. Giudice, G., Sanchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxf.)* **2016**, baw035 (2016).
28. Benoit Bouvrette, L. P., Bovaird, S., Blanchette, M. & Lecuyer, E. oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic Acids Res.* **48**, D166–D173 (2020).
29. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
30. Ghanbari, M. & Ohler, U. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res.* **30**, 214–226 (2020).
31. Gu, Z. et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat. Genet.* **51**, 296–307 (2019).
32. Spellman, R., Llorian, M. & Smith, C. W. Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1. *Mol. Cell* **27**, 420–434 (2007).
33. Spellman, R. & Smith, C. W. Novel modes of splicing repression by PTB. *Trends Biochem. Sci.* **31**, 73–76 (2006).
34. Haberman, N. et al. Insights into the design and interpretation of iCLIP experiments. *Genome Biol.* **18**, 7 (2017).
35. Buchbender, A. et al. Improved library preparation with the new iCLIP2 protocol. *Methods* **178**, 33–48 (2020).
36. Mikl, M., Hamburg, A., Pilpel, Y. & Segal, E. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. *Nat. Commun.* **10**, 4572 (2019).
37. Cheng, J. et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
38. Mount, S. M. et al. Assessing predictions of the impact of variants on splicing in CAGI5. *Hum. Mutat.* **40**, 1215–1224 (2019).
39. Yu, Y. et al. Dynamic regulation of alternative splicing by silencers that modulate 5′ splice site competition. *Cell* **135**, 1224–1236 (2008).
40. Julien, P., Miñana, B., Baeza-Centurion, P., Valcárcel, J. & Lehner, B. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. *Nat. Commun.* **7**, 11558 (2016).
41. Ledererova, A. et al. Hypermethylation of CD19 promoter enables antigen-negative escape to CART-19 in vivo and in vitro. *J. Immunother. Cancer* **9**, e002352 (2021).
42. Yoshida, K. et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).
43. Quesada, V. et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 47–52 (2011).
44. Black, K. L. et al. Aberrant splicing in B-cell acute lymphoblastic leukemia. *Nucleic Acids Res.* **46**, 11357–11369 (2018).
45. Desterro, J., Bak-Gordon, P. & Carmo-Fonseca, M. Targeting mRNA processing as an anticancer strategy. *Nat. Rev. Drug Discov.* **19**, 112–129 (2020).
46. Xu, Y. et al. Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing. *Genes Dev.* **28**, 1191–1203 (2014).
47. Itskovich, S. S. et al. MBNL1 regulates essential alternative RNA splicing patterns in MLL-rearranged leukemia. *Nat. Commun.* **11**, 2369 (2020).
48. Calabretta, S. et al. Modulation of PKM alternative splicing by PTBP1 promotes gemcitabine resistance in pancreatic cancer cells. *Oncogene* **35**, 2031–2039 (2016).
49. Monzón-Casanova, E. et al. Polypyrimidine tract-binding proteins are essential for B cell development. *Elife* **9**, e53557 (2020).
50. Shinohara, H. et al. Perturbation of energy metabolism by fatty-acid derivative AIC-47 and imatinib in BCR-ABL-harboring leukemic cells. *Cancer Lett.* **371**, 1–11 (2016).
51. Yap, K. et al. A short tandem repeat-enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. *Mol. Cell* **72**, 525–540 e513 (2018).
52. Shalabi, H. et al. Sequential loss of tumor surface antigens following chimeric antigen receptor T-cell therapies in diffuse large B-cell lymphoma. *Haematologica* **103**, e215–e218 (2018).
53. Spiegel, J. Y. et al. CAR T cells with dual targeting of CD19 and CD22 in adult patients with recurrent or refractory B cell malignancies: a phase 1 trial. *Nat. Med.* **27**, 1419–1431 (2021).
54. Venables, J. P. et al. Identification of alternative splicing markers for breast cancer. *Cancer Res.* **68**, 9525–9531 (2008).
55. Sheykhhasan M., Manoochehri H. & Dama P. Use of CAR T-cell for acute lymphoblastic leukemia (ALL) treatment: a review study. *Cancer. Gene Ther.* https://doi.org/10.1038/s41417-41021-00418-41411 (2022).
56. Fellmann, C. et al. Functional identification of optimized RNAi triggers using a massively parallel sensor assay. *Mol. Cell* **41**, 733–746 (2011).
57. Williams, R. et al. Amplification of complex gene libraries by emulsion PCR. *Nat. Methods* **3**, 545–550 (2006).
58. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
59. Estefania M., Andres R., Javier I., Marcelo Y. & Ariel C. ASpli: integrative analysis of splicing landscapes through RNA-Seq assays. *Bioinformatics*, btab141 (2021).
60. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

61. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
62. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
63. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
64. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
65. Schulz, L. et al. Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. *Genome Biol.* **22**, 190 (2021).
66. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
67. Pedregosa, F. et al. scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
68. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
69. Roehr, J. T., Dieterich, C. & Reinert, K. Flexbar 3.0 - SIMD and multicore parallelization. *Bioinformatics* **33**, 2941–2942 (2017).
70. Busch, A., Brüggemann, M., Ebersberger, S. & Zarnack, K. iCLIP data analysis: a complete pipeline from sequencing reads to RBP binding sites. *Methods* **178**, 49–62 (2020).
71. Krueger F. TrimGalore. *GitHub repository* (2021).
72. Vaquero-Garcia, J. et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**, e11752 (2016).
73. Howe, K. L. et al. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
74. Fernández, J. M. et al. The BLUEPRINT data analysis portal. *Cell Syst.* **3**, 491–495 e495 (2016).
75. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

Acknowledgements

The authors would like to thank the members of the participating labs for their support and discussion. We would like to thank Sylvia Weiss, Dpt. Systems Biology, University of Stuttgart for technical assistance. We gratefully acknowledge the Institute of Molecular Biology Core Facilities for their support, especially the Genomics Core Facility and the use of its NextSeq 500 (funded by the Deutsche Forschungsgemeinschaft [DFG, German Research Foundation] INST 247/870-1 FUGG) and the Bioinformatics Core Facility. We gratefully acknowledge the PacBio SMRT sequencing platform at MPI-CBG Dresden. TARGET data used for the analyses were accessed under dbGaP project #10088: "Alternative splicing in pediatric cancers", sub-studies phs000463.v21.p8 and phs000464.v21.p8 (Acute Lymphoblastic Leukemia (ALL) Pilot Phase 1 and Expansion Phase 2). This work was funded by the Naturwissenschaftlich-Medizinische Forschungszentrum (NMFZ) to J.F., J. K. and C. P. and the Deutsche Forschungsgemeinschaft (DFG) to K.Z., J.K. and S. L. (ZA 881/2–3 to K. Z., KO 4566/4-3 to J. K., and LE 3473/2–3 to S. L.). K. Z. was also supported by the Deutsche Forschungsgemeinschaft (SFB902 B13). This work was supported by the grant from the National Institutes of Health (U01 CA232563 to A. T.-T. and Y. B.), St. Baldrick's-Stand Up to Cancer (SU2C-AACR-DT-27-17 to A. T.-T.) and the V Foundation for Cancer Research (T2018-014 to A. T.-T.). M. T. D. acknowledges support from The Ellen Weisberg Fund: Advancing Breakthroughs in Pediatric Cancer.

Author contributions

M.C.-L. performed most bioinformatics analyses. L. S. performed the *CD19* minigene experiments as well as the massively parallel *CD19* splicing reporter assay. L.S. and B.S performed shRNA-mediated RBP knockdown experiments and corresponding splicing assays. M.M. and M.M.M. helped with experiments. M.E. and S.L. designed the mathematical modelling and prevalence score approach, and M.E. performed the analyses. F.K. contributed to the quantification of mutation effects. S.U. and M.K. validated single-mutation effects from model predictions. A.O., M.C.-L., L.S. and J.K. performed PTB iCLIP experiments. M.T.D. performed CD19 flow cytometry assays upon *PTBP1* knockdown and associated measurements. A.B. performed iCLIP and RNA-seq data processing as well as splice isoform quantification. M.Q.-V., M.T.D. and R.S. performed TARGET ALL data analysis under supervision of Y.B. and A.T.-T.. Study was designed by M.C.-L., L.S., M.E., K.Z., S.L. and J.K. with help from C.P., J.F. and all co-authors. K.Z., S.L. and J.K. supervised most of the bioinformatics analyses, mathematical modelling, and experimental work, respectively. M.C.-L., L.S., M.E., C.P., K.Z., S.L., and J.K. wrote the manuscript with help and comments from all co-authors.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

A.T.-T. has an interest in intellectual property "Discovery of CD19 Spliced Isoforms Resistant to CART-19". This interest does not meet the definition of a reviewable interest under Children's Hospital of Philadelphia's (CHOP's) conflict of interest policy and is therefore not a financial conflict of interest. Furthermore, this intellectual property has not been licensed or otherwise commercialised to date. However, should this technology be commercialised in the future, A.T.-T. would be entitled to a share of royalties earned by CHOP per its patent policy. The other authors have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-31818-y.

Correspondence and requests for materials should be addressed to Kathi Zarnack, Stefan Legewie or Julian König.

Peer review information *Nature Communications* thanks Dominique Payet-Bornet, Chonghui Cheng, Tim Mercer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

High-throughput mutagenesis identifies mutations and RNA-binding proteins controlling *CD19* splicing and CART-19 therapy resistance

Mariela Cortés-López[1#], Laura Schulz[1#], Mihaela Enculescu[1#], Claudia Paret[2,3,4], Bea Spiekermann[1], Mathieu Quesnel-Vallières[5,6], Manuel Torres-Diz[7], Sebastian Unic[8], Anke Busch[1], Anna Orekhova[1], Monika Kuban[5], Mikhail Mesitov[1], Miriam M. Mulorz[1], Rawan Shraim[7,9], Fridolin Kielisch[1], Jörg Faber[2,3,4], Yoseph Barash[5,6], Andrei Thomas-Tikhonenko[7,10], Kathi Zarnack[11,12,*], Stefan Legewie[1,8,13,*], and Julian König[1,*]

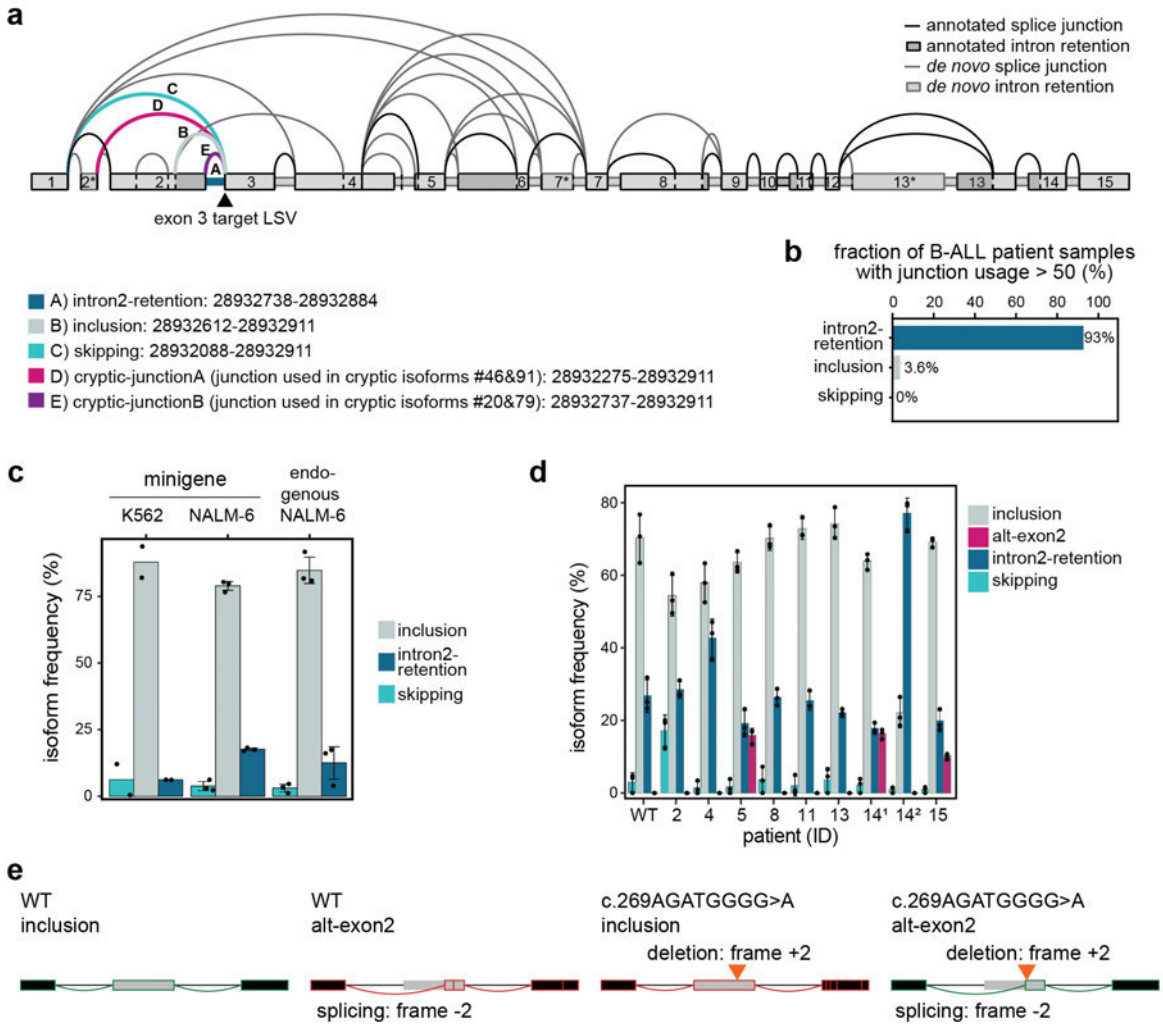
[1] Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany
[2] Department of Pediatric Hematology/Oncology, Center for Pediatric and Adolescent Medicine, University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany
[3] University Cancer Center (UCT), University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany
[4] German Cancer Consortium (DKTK), site Frankfurt/Mainz, Germany, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[5] Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
[6] Department of Biochemistry and Biophysics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
[7] Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
[8] Department of Systems Biology, Institute for Biomedical Genetics (IBMG), University of Stuttgart, Allmandring 30E, 70569 Stuttgart, Germany
[9] Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA
[10] Department of Pathology & Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
[11] Buchmann Institute for Molecular Life Sciences (BMLS), Max-von-Laue-Str. 15, 60438 Frankfurt, Germany
[12] Faculty Biological Sciences, Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt, Germany
[13] Stuttgart Research Center for Systems Biology (SRCSB), University of Stuttgart, Stuttgart, Germany

# These authors contributed equally.
* Corresponding authors: Kathi Zarnack (kathi.zarnack@bmls.de), Stefan Legewie (legewie@iig.uni-stuttgart.de), Julian König (j.koenig@imb-mainz.de)

SUPPLEMENTARY INFORMATION

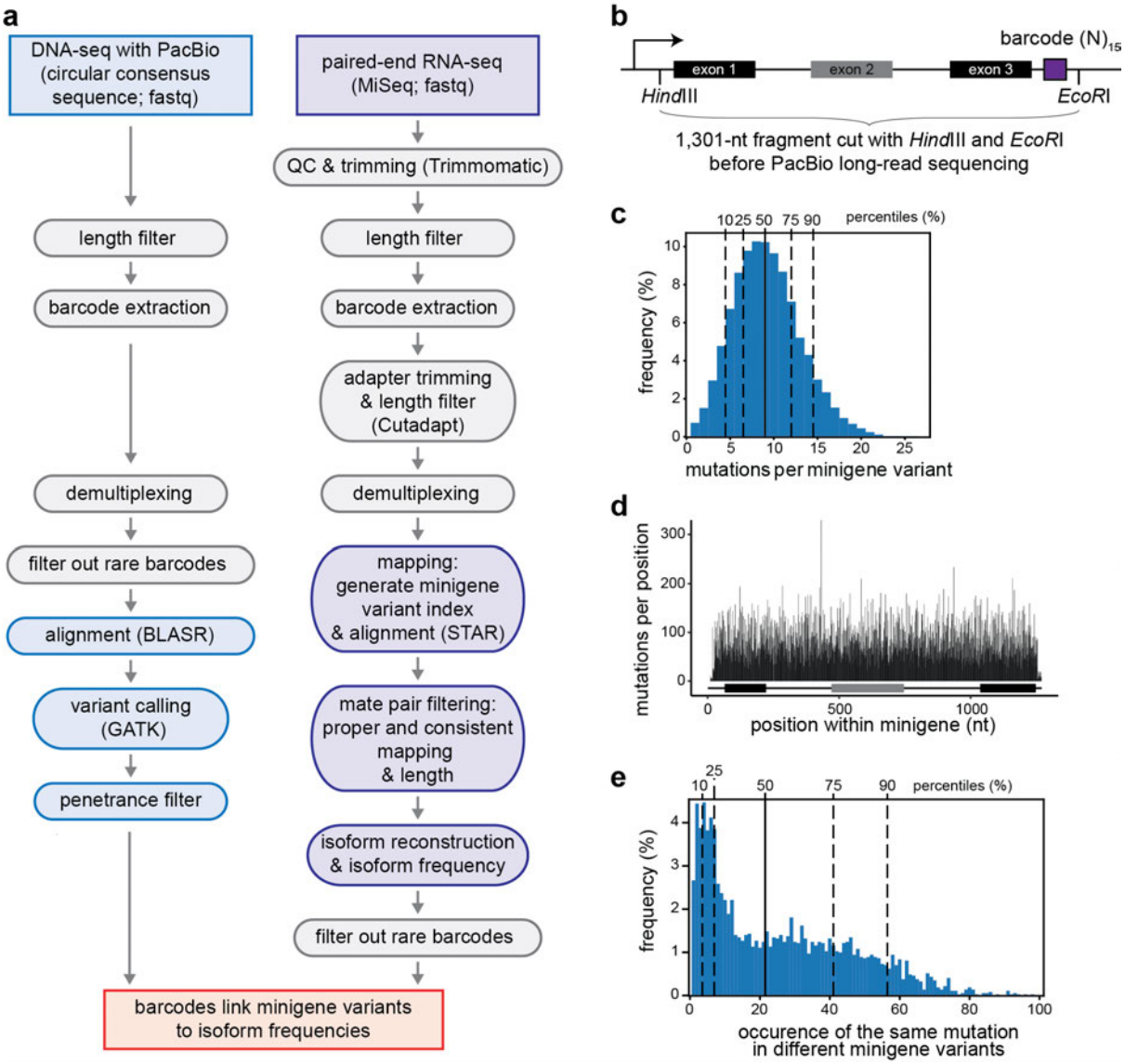
Table of content:

Supplementary Figures



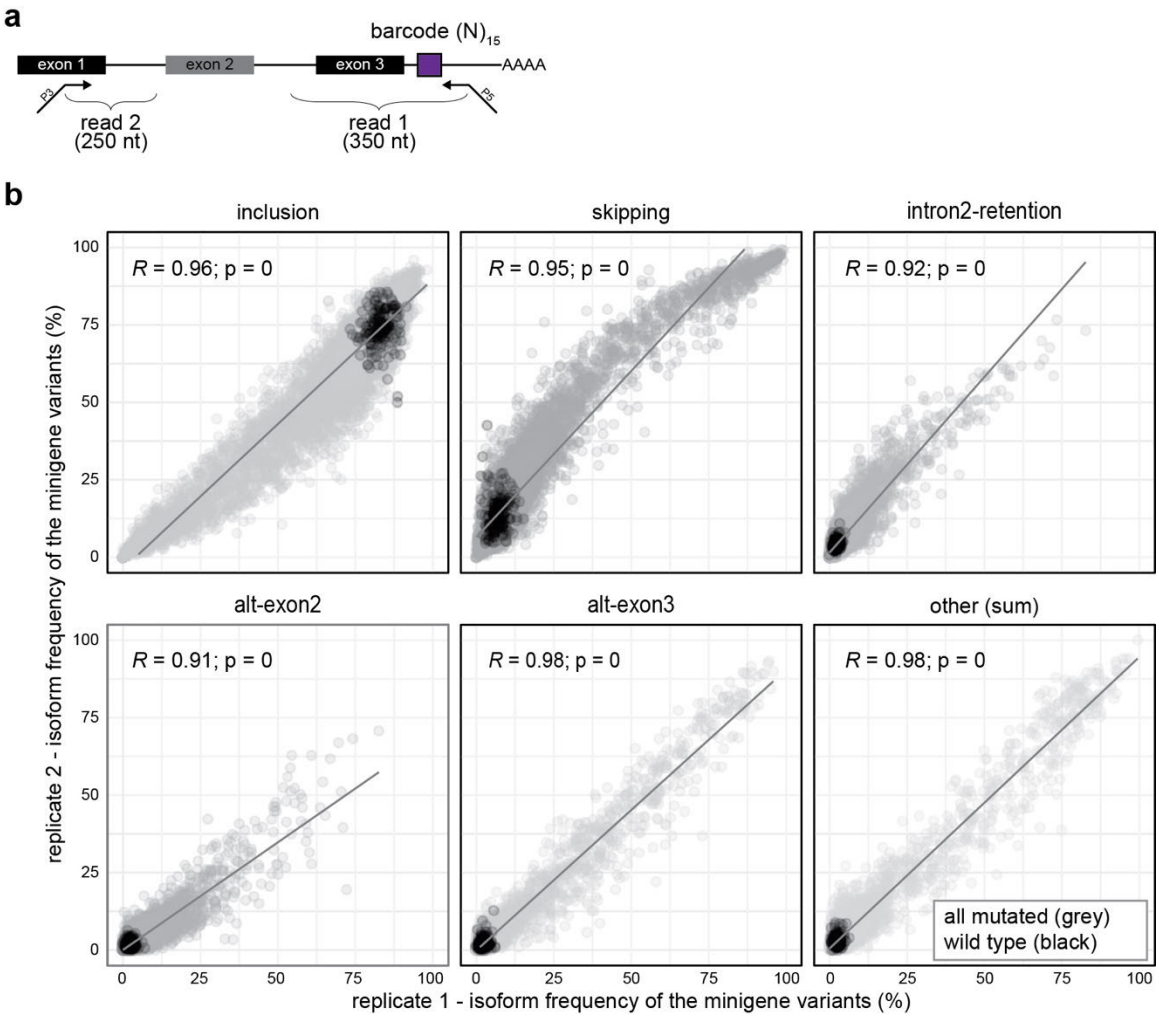
Supplementary Figure 1. *CD19* mis-splicing in TARGET B-ALL and Orlando datasets. **(a)** *CD19* shows extensive mis-splicing in B-ALL patients. Splice junctions were quantified with MAJIQ[1] for 220 B-ALL patients from the Therapeutically Applicable Research To Generate Effective Treatments (TARGET) programme (https://ocg.cancer.gov/programs/target). Splicegraph shows all splice junctions with a usage level (percent selected index, PSI) of at least 5% in any patient. Junctions and target exon of the local splicing variation (LSV) shown in (b) and Figure 1c, d are highlighted. **(b)** Intron 2 retention is the predominant isoform in B-ALL patients. Barchart quantifies the fraction of patients (220 B-ALL patients from the TARGET B-ALL programme) in which a given junction rises to PSI > 50%. **(c)** The minigene generates the same isoforms as the endogenous *CD19* gene in NALM-6 cells. Semiquantitative RT-PCR was performed to detect isoforms generated from exons 1-3 of the *CD19* minigene and the endogenous *CD19* gene in NALM-6 cells. Quantifications (mean and data points) of individual isoforms corresponding to Figure 1g. Error bars indicate standard deviation of mean (s.d.m.) if n > 2 replicates. **(d)** Patient mutations cause splicing changes in the *CD19* minigene. Semiquantitative RT-PCR as in (c) for minigene variants including nine mutations from B-ALL patients. Quantifications (mean and data points) of individual isoforms corresponding to Figure 1i. Patient ID numbers as reported in Orlando et al.[2]. 14.1 and 14.2 correspond to distinct mutations from patient #14. Error bars indicate s.d.m., n = 3 replicates. **(e)** The deletion c.269AGATGGGG>A from patient #5 in Orlando et al.[2] introduces a frameshift (+2) that is compensated by the activation of an out-of-frame cryptic splice site (-2). Shown are the major isoforms inclusion and alt-exon2 and their coding potential in the absence (left)

or presence (right) of the deletion (orange arrowhead). Schematic representation of depicts exons 1-3 (boxes) and introns (horizontal lines) with splice junctions for each isoform (arches). Colour indicates coding potential (green, coding; red, non-coding).

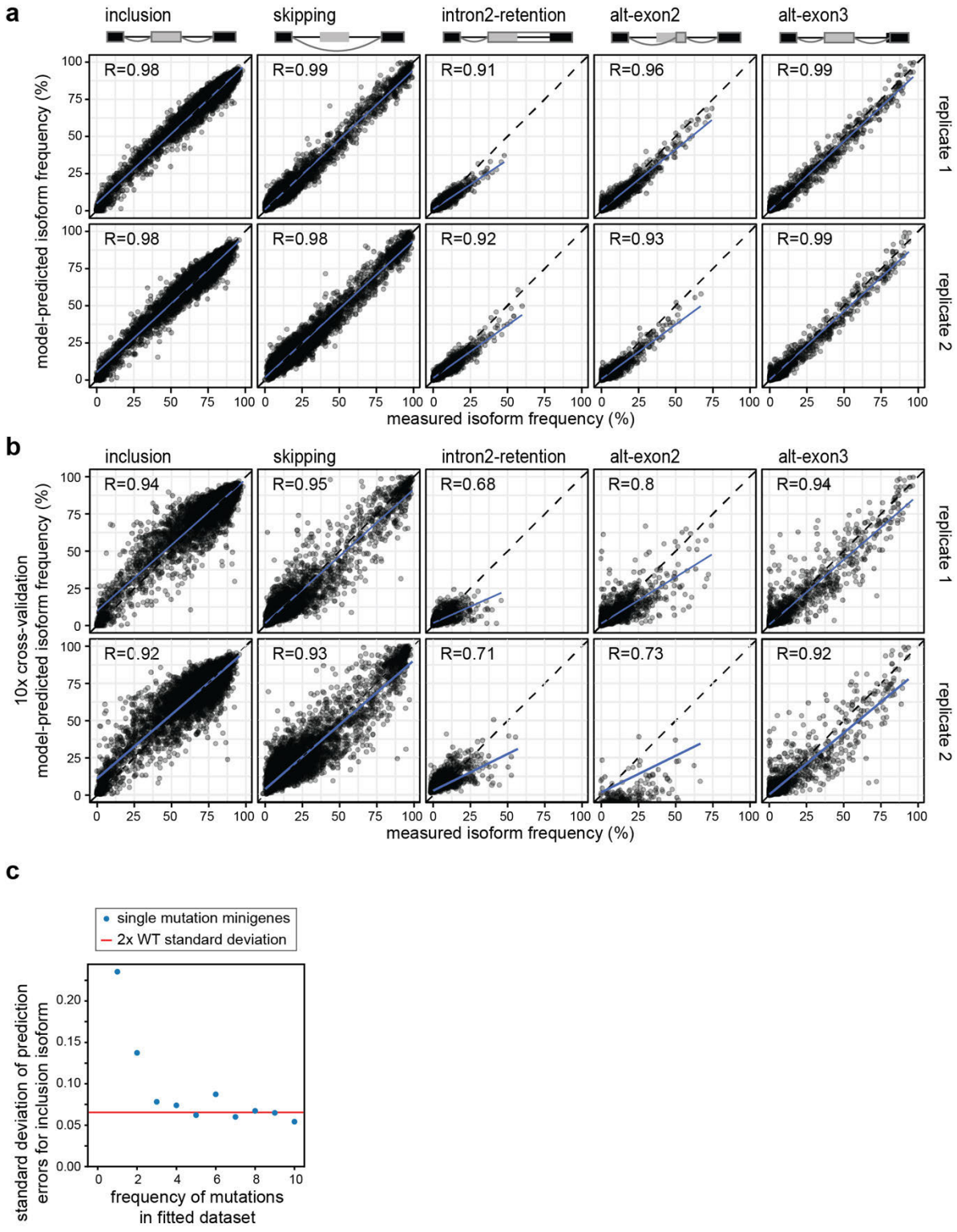


Supplementary Figure 2. Long-read sequencing identifies the introduced mutations. **(a)** Analysis pipeline for the targeted DNA-seq and RNA-seq data. Left: Long-read DNA-seq data (PacBio, Pacific Bioscience) in the form of circular consensus sequences (CSS) were filtered by length (1,150-1,500 nt). 15-nt barcodes were extracted and demultiplexed, keeping only minigenes supported by at least 4 CSS. Alignment to the minigene reference was performed with BLASR[3] and variants were called using GATK HaplotypeCaller[4]. Mutations in the minigene were filtered by the "penetrance score" (allele frequency, AF), discarding all the barcodes with more than 25% variants of low penetrance (AF < 0.8). Right: Short-read RNA-seq data (Illumina) were trimmed based on quality using Trimmomatic[5] and filtered by length (305 nt for read 1, 157 nt for read 2), and adapters were trimmed using Cutadapt[6] and 15-nt barcodes were extracted and demultiplexed, keeping only minigenes supported by at least 100 read pairs. Alignment to the specific mutated version of the minigene was performed using STAR[7]. Isoform reconstruction and isoform frequency estimation was done using custom scripts (see Methods). Only minigenes with 100 or more read pairs usable for isoform reconstruction were kept. **(b)** Structure of the *CD19* minigene fragment for long-read sequencing (PacBio) to identify introduced mutations. The minigene covers exons 1-3 with the intervening introns, followed by a 15-nt barcode. The fragment for PacBio sequencing is defined by the restriction sites for *Hind*III upstream of exon 1 and *EcoR*I downstream of the barcode sequence. **(c)** 91.6% of the minigene variants carry five or more mutations. Histogram shows number of mutations per minigene for 10,295 mutated minigene variants. **(d)** 4,255

distinct mutations are spread along the *CD19* minigene, with an average of 21 mutations per position. Barplot shows the sum of mutations per position in the minigene. **(e)** 81.9% of the mutations occur in at least three minigenes, which is sufficient for a reliable estimation of single mutation effects (Supplementary Figure 4c). Histogram shows the frequencies of the same mutations in different minigene variants.

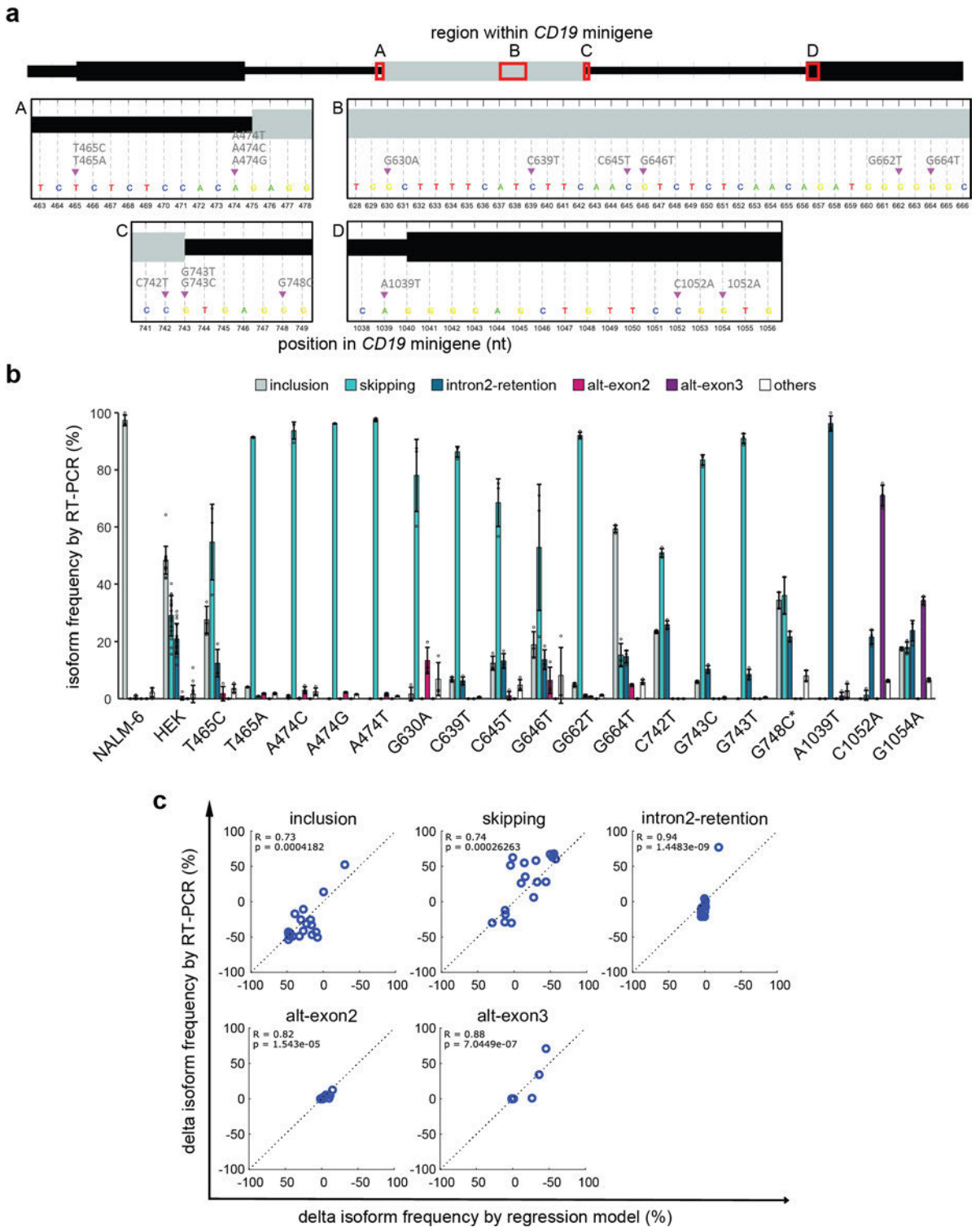


Supplementary Figure 3. Isoform measurements from targeted RNA-seq results are consistent between replicates. (a) Description of the short-read RNA-seq strategy (Illumina) to capture the splicing products in the *CD19* minigene. Read 2 (250 nt) extends beyond exon 1, i.e., covering the exon 1/exon 2 junction, while read 1 (350 nt) includes the 15-nt barcode and extends beyond exon 3. **(b)** The isoform measurements correlate well between replicates. Scatterplots compare isoform frequencies for five major isoforms as well as the sum of 96 cryptic isoforms between replicate 1 and 2. Each dot represents a particular minigene captured in both replicates. WT and mutated minigenes appear in black and grey, respectively. Pearson correlation coefficients (*R*) and associated *P* values (two-sided) are given.



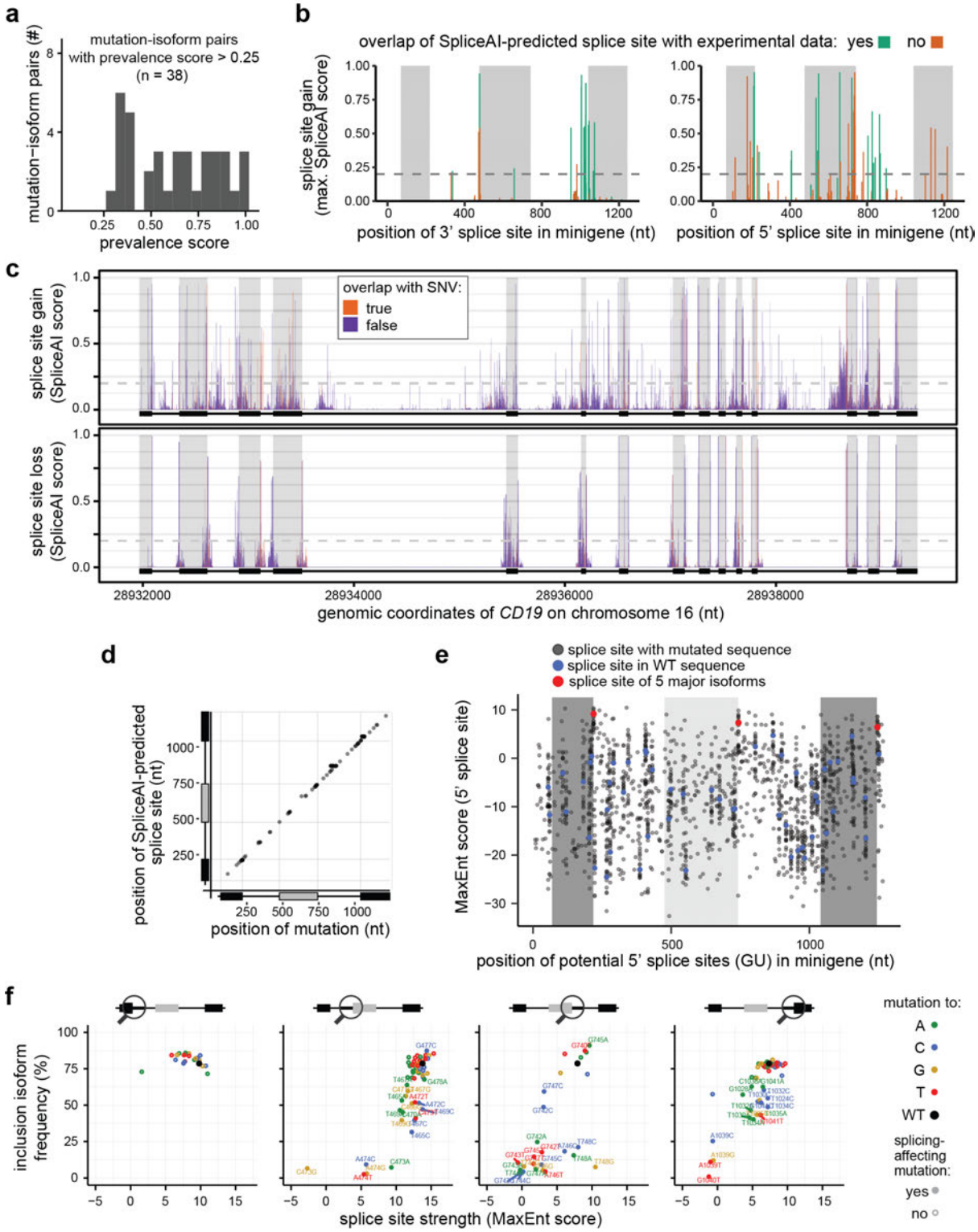
Supplementary Figure 4. The softmax regression model performs well for training and test data. (a) Regression model fits measured combined mutation effects (i.e., minigene measurements) with high accuracy. Scatterplots show frequencies of the five major isoforms in the measurements (x-axis) against the model fit (y-axis) for two biological replicates and 9,321 minigene variants used in model training. Pearson correlation coefficients (*R*) are shown for each scatterplot. **(b)** Cross-validation confirms the predictive power of the model for minigenes not used in training. The minigene library was randomly split into ten equally sized subsets. During 10-fold cross-validation, the softmax regression model was fitted to all data

excluding one subset. Scatterplots compare model-predicted splicing outcome for left-out subsets to corresponding experimental data for all major splice isoforms and are an overlay of the results of all cross-validation runs. Representation as in (a). **(c)** The model correctly infers single mutation effects. Seven single-mutation minigenes in which inclusion is significantly changed were left-out separately from softmax regression fitting and their effects were predicted based on the fit to the remaining minigene data. This procedure was repeated while additionally excluding random permutations of other minigenes containing the mutation. The standard deviation of the prediction error (y-axis) is plotted against the number of minigenes used in model training (x-axis). The inference power of the model reaches two standard deviations of the WT minigenes (horizontal line) if more than two minigenes containing the mutation are considered in model training. See Methods for details.



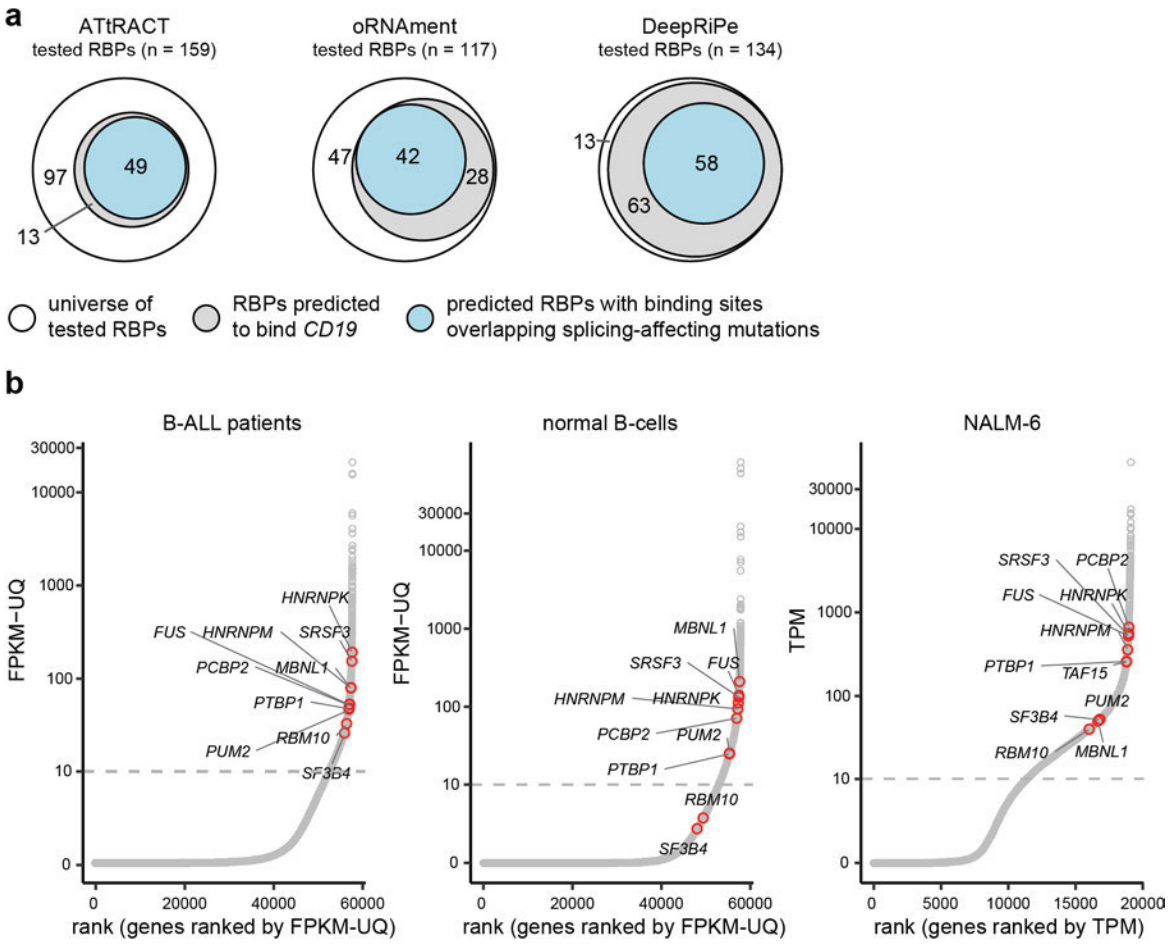
Supplementary Figure 5. RT-PCR measurements confirm the model predictions for 19 individual point mutations. (a) To test selected regression predictions, we generated 19 minigenes with individual point mutations that are predicted to affect at least one isoform (Supplementary Data 4). Point mutations were introduced by targeted mutagenesis. **(b)** Splicing outcome was quantified using RT-PCR followed by capillary electrophoresis. Quantifications (mean and data points) of individual isoforms corresponding to Figure 3e. 'NALM-6', splicing pattern of WT minigenes (RNA-seq) in the mutagenesis screen, 'HEK293', RT-PCR-based quantification of the baseline minigene containing mutation G742C in HEK293 cells. G748C* is a minigene containing G748C but lacking G742C. Error bars indicate

standard s.d.m. if n > 2 replicates. **(c)** Splicing patterns in response to single mutations correlate with regression predictions. Splicing outcomes from 19 *CD19* minigene variants containing single point mutations (y-axis) are related to single mutation predictions of the regression model (x-axis; mean of two fits, each explaining one mutagenesis replicate). Changes in the isoform frequency of the major isoforms are expressed as differences (delta) relative to the baseline. Pearson correlation coefficients and *P* values (two-sided) were calculated for each isoform (see Figure 3f for correlation over all isoforms).

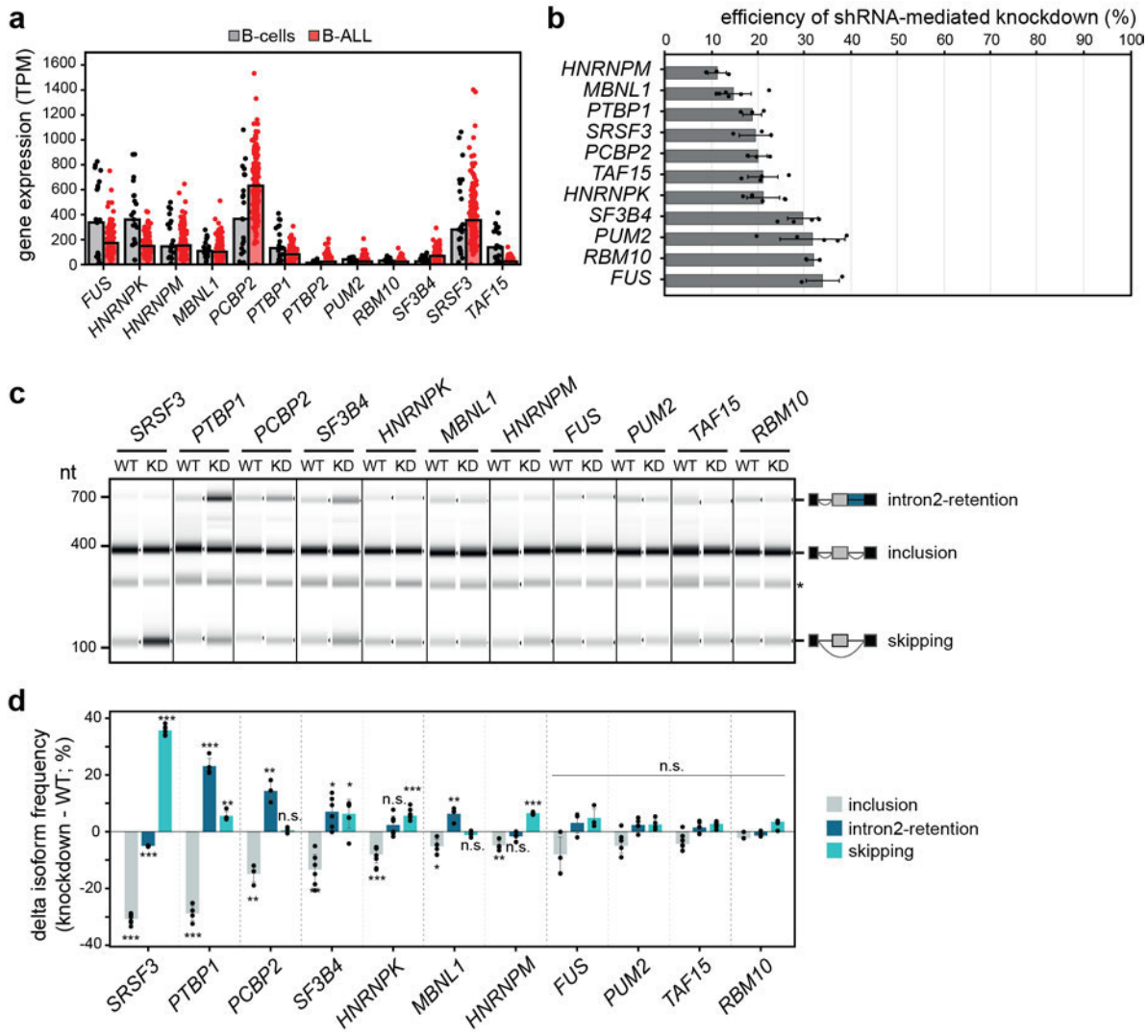


Supplementary Figure 6. Multiple mutations give rise to distinct cryptic isoforms. **(a)** Multiple mutations are associated with a specific cryptic isoform. Histogram shows distribution of prevalence scores for 38 mutation-isoform pairs (prevalence score > 0.25). A prevalence score of 1 indicates perfect correspondence between mutation and isoform. **(b)** SpliceAI[8] predictions for gained cryptic splice sites overlap with experimental data. Barplot shows the maximum SpliceAI score ("acceptor gain") for all the mutations that increase the probability of a given cryptic splice site to be used (38 mutations with Splice AI score [gain] > 0.5, including 15 and 23 gained 3' [left] and 5' splice sites [right]). Dotted horizontal line represents the recommended minimum threshold for a SpliceAI prediction (SpliceAI score >

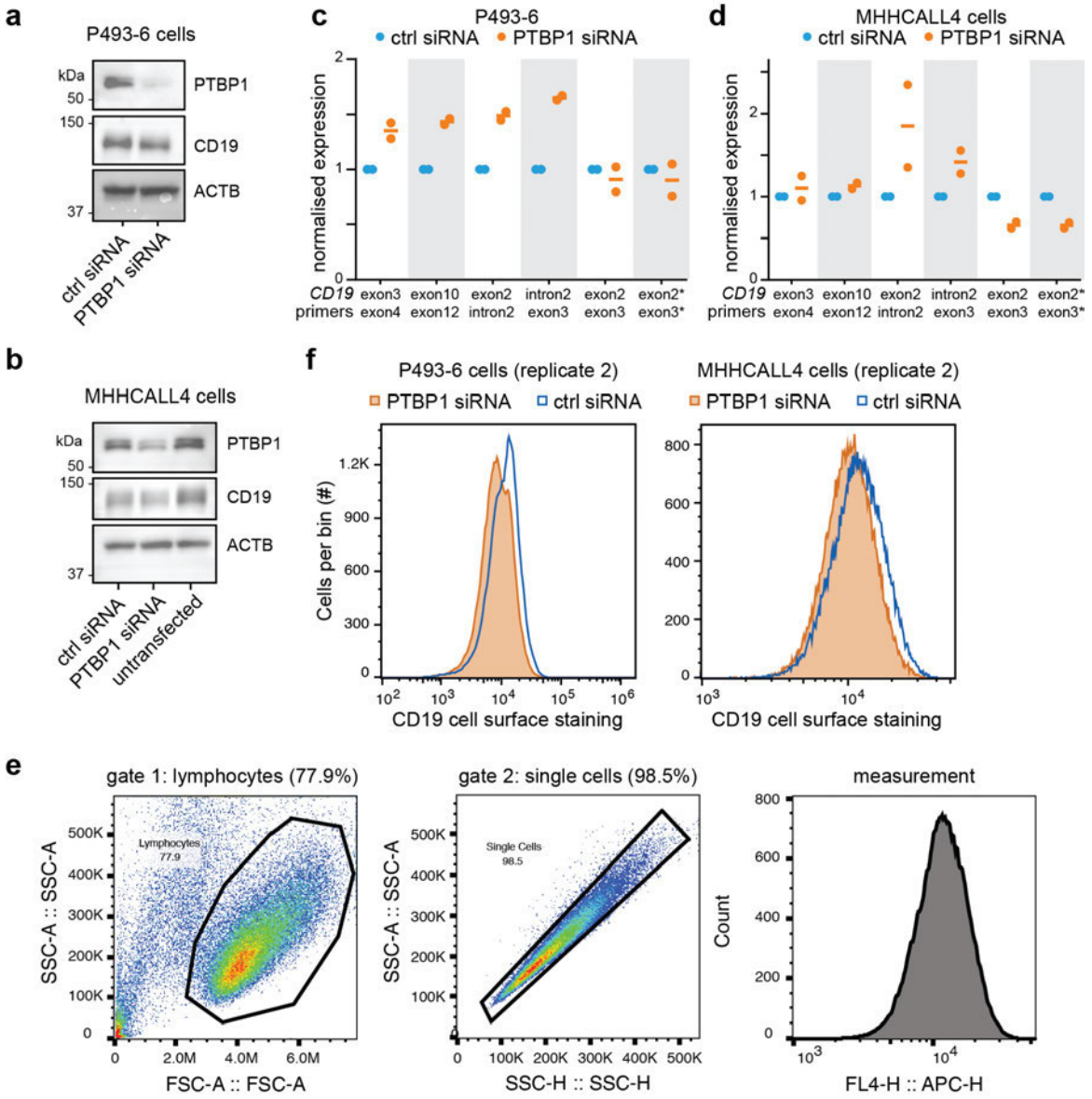
0.2)[8]. Predicted gained splice sites that also appear in our experimental data are shown in green. **(c)** SpliceAI predicts splice-changing mutations across the full *CD19* gene locus. Barplots show the maximum SpliceAI score per position. Scores are separately shown for the gain (top) or loss (bottom) of splice sites. Colour code indicates overlap with reported variants (from gnomAD, ClinVar, COSMIC V94, Ensembl and TARGET B-ALL). 24 and 13 mutations reach a SpliceAI score > 0.2 for the gain and loss of splice sites, respectively (Supplementary Data 5). **(d)** SpliceAI-predicted splicing-affecting mutations reside on average within 6 nt from the cryptic splice site generated. Scatterplot shows location of the gained cryptic splice sites with respect to the mutations. Only the splice site with the highest score for each mutation is considered. **(e)** The 5' splice sites of the main isoforms (red) are stronger than most other 5' splice sites in the *CD19* minigene sequence. Dotplot shows splice site strengths (MaxEnt score)[9] for putative 5' splice sites in WT (blue) and mutated (grey) minigenes in a 9-nt sliding window containing a GU dinucleotide at positions 4-5. 5' splice sites used in the five major isoforms are shown in red. **(f)** Mutation effects at 3' and 5' splice sites of *CD19* exons 2 and 3 are consistent with predicted splice site strengths. Mutations are coloured according to the changed nucleotides. Scores for WT sequence are coloured in black. Splicing-affecting mutations (according to our results) are shown as filled circles and labelled.



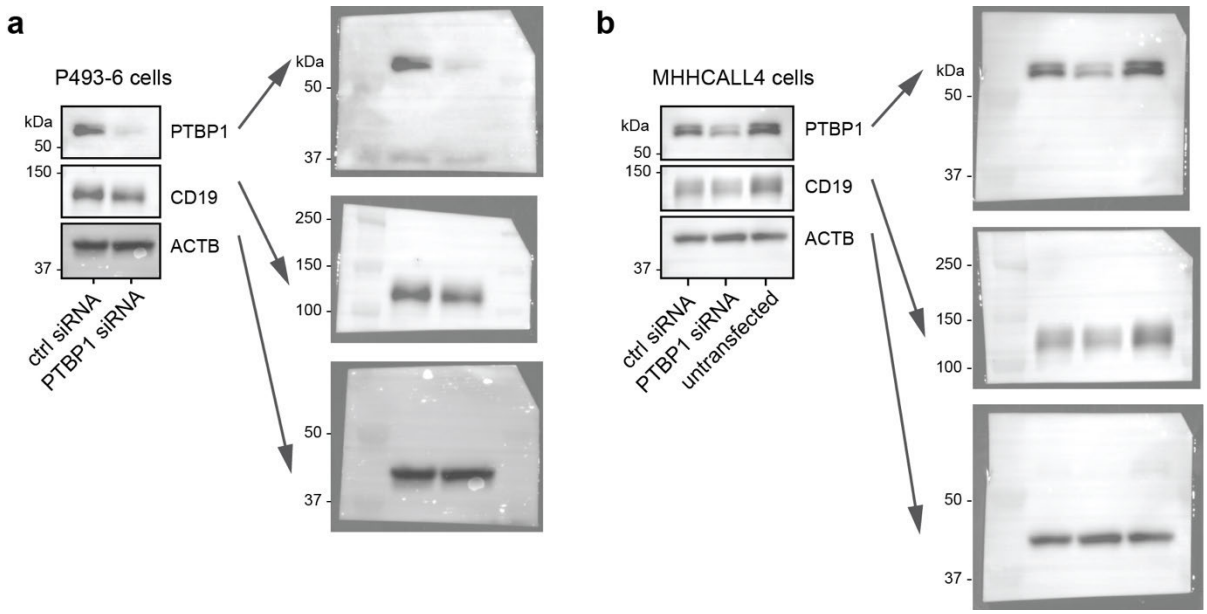
Supplementary Figure 7. *In silico* RBP binding site predictions suggest dozens of candidate regulators of *CD19* alternative splicing. (a) *In silico* predictions of RBP binding sites were performed with ATtRACT[10] and oRNAment[11] as well as of point mutations affecting RBP binding using DeepRiPe[12]. For each prediction tool, the total number of available RBPs (white circles) is split up into those that are predicted to bind *CD19* (grey circles) and whose predicted binding sites overlap with splicing-affecting mutations from our data (blue circles). Numbers refer to exclusive RBPs in each area. **(b)** Predicted RBPs were filtered based on their mean expression observed in B-ALL patients reported in[13]. Plot shows ranked mean expression values for all detected genes in samples from B-ALL patients (n = 57,773 genes, 1,988 patients), normal B-cells[14] (n = 57,773 genes, 147 samples) and NALM-6 cells[15] (n = 19,110 genes, 1 sample). Highlighted in red are the RBP candidate genes (n = 11) tested in knockdown experiments. TPM, transcripts per million. FPKM-UQ, fragments per kilobase of transcript per million mapped reads upper quartile, a modified RNA-seq normalisation method (https://docs.gdc.cancer.gov/Encyclopedia/pages/HTSeq-FPKM-UQ/).



Supplementary Figure 8. Knockdown experiments show significant effects on endogenous *CD19* splicing for seven candidate RBPs. (a) The tested RBPs are expressed in patients. Barplot shows RBP mRNA levels (TPM) for normal B-cells (n = 21) and TARGET B-ALL patient samples (n = 220). **(b)** All tested RBPs are efficiently depleted upon shRNA knockdown (KD). Barplot shows mean qPCR measurements of remaining transcripts (relative to WT) for 11 candidate RBPs. Error bars indicate standard deviation of the mean (s.d.m.), n = 3 replicates. **(c, d)** Seven RBP knockdowns significantly affect *CD19* alternative splicing. Semiquantitative RT-PCR was performed to detect isoforms generated from exons 1-3 of the endogenous *CD19* gene. Gel-like representation (c), with major isoforms indicated on the right, and quantification (d), as difference in isoform frequency compared to WT, are shown. Error bars indicate s.d.m., n = 3 replicates. * *P* value < 0.05, ** *P* value < 0.01, *** *P* value < 0.01, n.s., not significant, two-sided Student's *t*-test. Source data including *P* values are provided as a Source Data file.



Supplementary Figure 9. PTBP1 regulates CD19 protein surface expression. **(a, b)** Western blot analysis shows reduced PTBP1 and CD19 protein expression upon siRNA-mediated *PTBP1* knockdown in P493-6 (a) and MHHCALL4 (b) cells, two human B-cell lines derived from immortalised lymphocytes and B-ALL tumour cells, respectively (n = 2, exemplary data are shown). Actin B (ACTB) served as loading control. Uncropped images of the gels are provided in Supplementary Figure 10. **(c, d)** *CD19* intron 2 retention is increased upon *PTBP1* knockdown in P493-6 (c) and MHHCALL4 (d) cells. Barplots show qPCR quantification of different exon-exon and exon-intron junctions as indicated below. Samples were normalised to *GAPDH* mRNA and the non-targeting control siRNA condition. Error bars indicate standard deviation (n = 2 biological replicates). **(e)** Gating strategy for the flow cytometry analysis of CD19 surface protein exposure. The first gate was set for the cell population, the next gate for singlets and finally, immunostaining of CD19 surface protein was measured in the allophycocyanin (APC) channel. **(f)** CD19 cell surface staining is reduced upon *PTBP1* knockdown in P493-6 (left panel; replicate 2) and MHHCALL4 (right panel; replicate 2) cells. Distributions of CD19 surface protein, as measured in 45-50x10^3 cells per replicate by CD19 antibody staining and flow cytometry, in cells transfected with PTBP1 siRNA (orange) or non-targeting control siRNA (blue). The results for replicate 1 are shown in Figure 6d, e.



Supplementary Figure 10. Uncropped images for Western blots in Supplementary Figure 9a, b. Western blot analysis shows reduced PTBP1 and CD19 protein expression upon siRNA-mediated *PTBP1* knockdown in P493-6 and MHHCALL4 cells. Actin B (ACTB) served as loading control.

Supplementary Tables

Supplementary Table 1. Mutations from relapsed B-ALL patients reported in Orlando et al. that were tested in the *CD19* minigene splicing reporter. Patient IDs are given as reported in Orlando et al.[2]. Note that for patient #14, two separate minigene variants were tested (#14.1 and #14.2), and that #14.2 is a combination of two adjacent mutations reported in patient #14, namely c.509A>AGTGG and c.510GCCTC>GTGGGGGAG.

patient ID	mutation	genomic coordinate (hg38)	position in minigene	reference allele (REF)	alternative allele (ALT)
#2	c.259G>GGGGGC	chr16:28932516	646	G	GGGGGC
#4	c.517TGTCTCCCACCG>T	chr16:28933072	1202	TGTCTCCCACCG	T
#5	c.269AGATGGGG>A	chr16:28932526	656	AGATGGGG	A
#8	c.265CA>C	chr16:28932522	652	CA	C
#11	c.264TCAACAGATGGGGGGCTTCTACCTGTGC>T	chr16:28932521	651	TCAACAGATGGGGGGCTTCTACCTGTGC	T
#13	c.421T>TC	chr16:28932976	1106	T	TC
#14.1	c.297GGGGC>G	chr16:28932554	684	GGGGC	G
#14.2	c.510AGCCTC>AGTGGGGGAG	chr16:28933065	1195	AGCCTC	AGTGGGGGAG
#15	c.271ATGGGGGGCTTCTACCTGTGCCAGCCGGGGCCC>AAGACGT	chr16:28932528	658	ATGGGGGGCTTCTACCTGTGCCAGCCGGGGCCC	AAGACGT

Supplementary Table 2. Oligonucleotides used to clone the different shRNA sequence carrying vectors in this study. Oligonucleotides were purchased from Integrated DNA Technologies.

shRNA_FUS	TGCTGTTGACAGTGAGCGCACAGGATAATTCAGACAACAATAG TGAAGCCACAGATGTATTGTTGTCTGAATTATCCTGTTTGCCTA CTGCCTCGGA
shRNA_HNRNPK	TGCTGTTGACAGTGAGCGACGAGTTGAGGCTGTTGATTCATAG TGAAGCCACAGATGTATGAATCAACAGCCTCAACTCGCTGCCT ACTGCCTCGGA
shRNA_HNRNPM	TGCTGTTGACAGTGAGCGAAGCAGACATTCTTGAAGATAATAGT GAAGCCACAGATGTATTATCTTCAAGAATGTCTGCTCTGCCTAC TGCCTCGGA
shRNA_MBNL1	TGCTGTTGACAGTGAGCGCCAGCACAATGATTGACACCAATAG TGAAGCCACAGATGTATTGGTGTCAATCATTGTGCTGTTGCCTA CTGCCTCGGA
shRNA_PCBP2	TGCTGTTGACAGTGAGCGCTCCATCATTGAGTGTGTCAAATAGT GAAGCCACAGATGTATTTGACACACTCAATGATGGATTGCCTAC TGCCTCGGA
shRNA_PTBP1	TGCTGTTGACAGTGAGCGCTAGCAAGATGATACAATGGTATAG TGAAGCCACAGATGTATACCATTGTATCATCTTGCTATTGCCTA CTGCCTCGGA
shRNA_PUM2	TGCTGTTGACAGTGAGCGCAACATAGTTGTTGACTGTTAATAGT GAAGCCACAGATGTATTAACAGTCAACAACTATGTTATGCCTAC TGCCTCGGA
shRNA_RBM10	TGCTGTTGACAGTGAGCGCCGGCAAGACCATCAATGTTGATAG TGAAGCCACAGATGTATCAACATTGATGGTCTTGCCGTTGCCTA CTGCCTCGGA
shRNA_SF3B4	TGCTGTTGACAGTGAGCGCTGCCTTCAAGAAGGACTCCAATAG TGAAGCCACAGATGTATTGGAGTCCTTCTTGAAGGCATTGCCTA CTGCCTCGGA
shRNA_SRSF3	TGCTGTTGACAGTGAGCGCTAAGATGTTTTAGCTGTTCAATAGT GAAGCCACAGATGTATTGAACAGCTAAAACATCTTAATGCCTAC TGCCTCGGA
shRNA_TAF15	TGCTGTTGACAGTGAGCGATCAGGCTATGATCAACATCAATAGT GAAGCCACAGATGTATTGATGTTGATCATAGCCTGACTGCCTAC TGCCTCGGA

Supplementary Table 3. qPCR oligonucleotide pairs used in this study. Oligonucleotides were purchased from Sigma-Aldrich.

	Forward primer	Reverse primer
qPCR_FUS	AAGGCCTGGGTGAGAATGTT	GGCTGTCCCGTTTTCTTGTT
qPCR_HNRNPK	GCGAGTTGAGGCTGTTGATT	TCAGTGGAATGAGGACAGCA
qPCR_HNRNPM	GTCAAGGGGATGTGCTGTTG	TCCGCTCAGACTATGCTTGT
qPCR_MBNL1	CGGTTTGCTCATCCTGCTGA	TTTGCACTTTTCCCGAGAGC
qPCR_PCBP2	CCAGCTCTCCGGTCATCTTT	CTGGTGCAGCTTGGTCAAAT
qPCR_PTBP1	CGAGATGAACACGGAGGAGG	CTGGATGTAGATGGGCTGGC
qPCR_PUM2	TCAGCGTCCTCTTACTCCCA	CCAGTAGCAAGACCCTGACC
qPCR_RBM10	TGTTCCCGACGTCTCTACCT	TCTCCCCATCCCAGTACAGG
qPCR_SF3B4	GAACGACTTCTGGCAGCTCA	CACAGGATTGGGAGCAGAGG
qPCR_SRSF3	CCCGGCTTTGCTTTTGTTGA	TTCCACTCTTACACGGCAGC
qPCR_TAF15	GGTCACAGGGAGGAGGTAGA	CAGCATCTGTTCTGGGTCCA
qPCR_CD19_E3E4	TGAGATCTGGGAGGGAGAG	ATCGTCCTTCAGCTCTAGGC
qPCR_CD19_E10E12	TCCTTCTCCAACGCTGAGTC	GAAGTCCATTGTCCTGGCGA
qPCR_CD19_e2i2	TGGCTGGACAGTCAATGTG	TCTCTCCAGCTCCATTGTGG
qPCR_CD19_i2e3	TCAGTATGAGCTGCTTCCTGTCC	AGCTCCCCTGGGAAGAGACC
qPCR_CD19_E2E3_1	AGGCCTGGGAATCCACATGA	GGAACAGCTCCCCGCTG
qPCR_CD19_E2E3_2	AGTCCCCGCTTAAACCCTTC	AGCTCCCCGCTGCCC
qPCR_GAPDH	ATGGGGAAGGTGAAGGTCG	GGGGTCATTGATGGCAACAATA
qPCR_ACTB	AGCATCCCCCAAAGTTCAC	AAGGGACTTCCTGTAACAACG

Supplementary References

1. Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife* **5**, e11752 (2016).
2. Orlando, E. J. *et al.* Genetic mechanisms of target antigen loss in CAR19 therapy of acute lymphoblastic leukemia. *Nat Med* **24**, 1504-1506 (2018).
3. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
4. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
5. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
6. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).
7. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
8. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548 e524 (2019).
9. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377-394 (2004).
10. Giudice, G., Sanchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016** (2016).
11. Benoit Bouvrette, L. P., Bovaird, S., Blanchette, M. & Lecuyer, E. oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic Acids Res* **48**, D166-D173 (2020).
12. Ghanbari, M. & Ohler, U. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res* **30**, 214-226 (2020).
13. Gu, Z. *et al.* PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat Genet* **51**, 296-307 (2019).
14. Alexander, T. B. *et al.* The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* **562**, 373-379 (2018).
15. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).

4. FUBP1 is a general splicing factor facilitating 3' splice site recognition and splicing of long introns

4.1. Abstract

Splicing is an integral part of gene regulation and proteomic diversity. Numerous studies unraveled the biochemical mechanisms of splicing in detail. However, it is less understood how splicing splice site recognition and splicing regulation takes place. Here, we identify the DNA- and RNA-binding protein FUBP1 to be a core splicing protein. Using high-throughput sequencing technologies, we found that FUBP1 ubiquitously binds pre-mRNA upstream of the branch point to a hitherto unknown U-rich motif. NMR-based experiments and reporter assays revealed that the N-terminal N-box of FUBP1 can interact with the RRM2 domain of U2AF2 and thereby stabilizes the latter at the py-tract. The C-terminal A/B-boxes, an animal-specific domain in FUBP proteins, are capable to interact both with SF1 and proteins of the U1 snRNP at the 5' ss. We established a *FUBP1* KO and a *FUBP1* mutant line by gene editing, and upon RNA-sequencing these cells lines, we could show that FUBP1 is crucial for splicing of long introns. We hypothesize that FUBP1 is a core splicing factor that regulates splice site bridging of long introns by interacting with proteins of both the 3'ss and the 5'ss.

4.2. Zusammenfassung

Spleißen ist ein integraler Bestandteil der Genregulation und proteomischer Diversität. Zahlreiche Studien, haben die biochemischen Mechanismen im Detail entwirrt. Wie jedoch die Spleißstellen erkannt und Spleißregulation stattfindet, ist weniger erforscht. In dieser Studie identifizieren wir das DNA- und RNA-Bindeprotein FUBP1 als zentrales Spleißprotein. Mithilfe von Hochdurchsatzsequenzierungstechnologien konnten wir feststellen, dass FUBP1 ubiquitär an ein bisher unbekanntes U-reiches Motiv in der prä-RNA bindet. NMR-basierte Experimente und Reporterassays enthüllten, dass die N-terminale N-box von FUBP1 mit der RRM2-Domäne von U2AF2 interagieren kann und somit Letzteres am Py-Trakt stabilisiert. Die C-terminalen A/B-Boxen, tierspezifische Domänen in FUBP-Proteinen, sind in der Lage, sowohl mit SF1 als auch mit Proteinen des U1 snRNP an der 5' Spleißstelle zu interagieren. Wir etablierten eine *FUBP1*-KO-Zelllinie und eine *FUBP1*-Mutantenzelllinie. Mithilfe von RNA-Sequenzierung dieser Zelllinien konnten wir demonstrieren, dass FUBP1 entscheidend für das Spleißen langer Introns ist. Wir schlagen die Hypothese vor, dass FUBP1 ein zentrales Spleißprotein ist, dass die Überbrückung von Spleißstellen reguliert, indem es mit Proteinen der 3' Spleißstelle als auch der 5' Spleißstelle interagiert.

4.3. Statement of contribution

This was the main project shared between [redacted], [redacted] and me. [redacted] performed the bioinformatic part while [redacted] performed all NMR-related tasks. [redacted], together with others, performed the iCLIP experiments. I validated the CRISPR/Cas9 Knockouts and performed all experiments using these cell lines, including the RNA-seq, Western Blots, the *MPDZ* minigene splicing assay and the complementation experiments. For these experiments, I created *MPDZ* minigene mutants and truncated FUBP1 variants by site-directed mutagenesis.

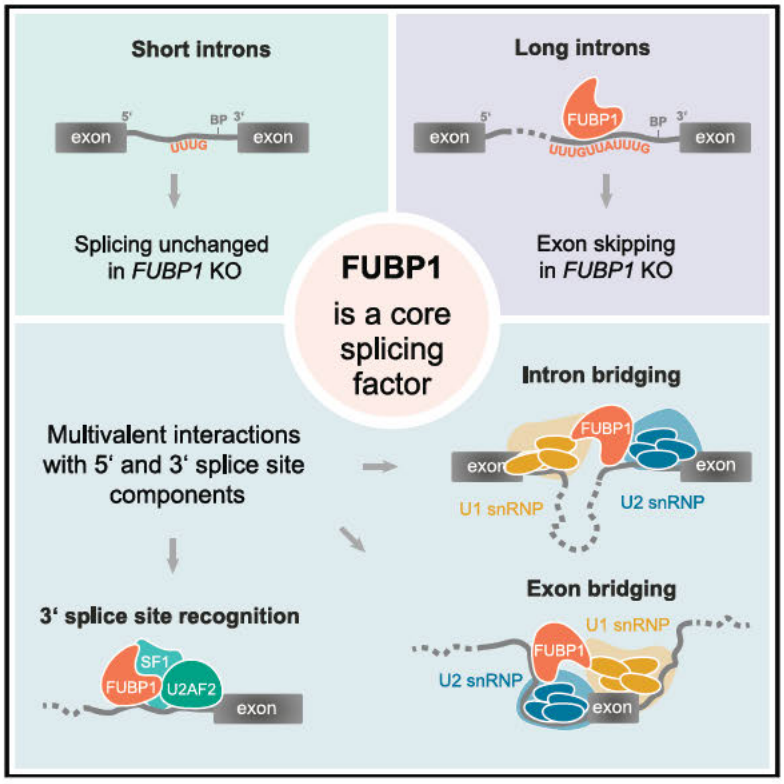
I took part in the design of the study, and in regular meetings with [redacted], but also with [redacted], I contributed to the discussion and interpretation of results. I prepared the figures corresponding to my experiments and wrote the corresponding results and methods. I arranged and edited all figures for the manuscript. Together with [redacted] and [redacted], I collected and organized material from collaborators, and we wrote, arranged and reviewed the manuscript.

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

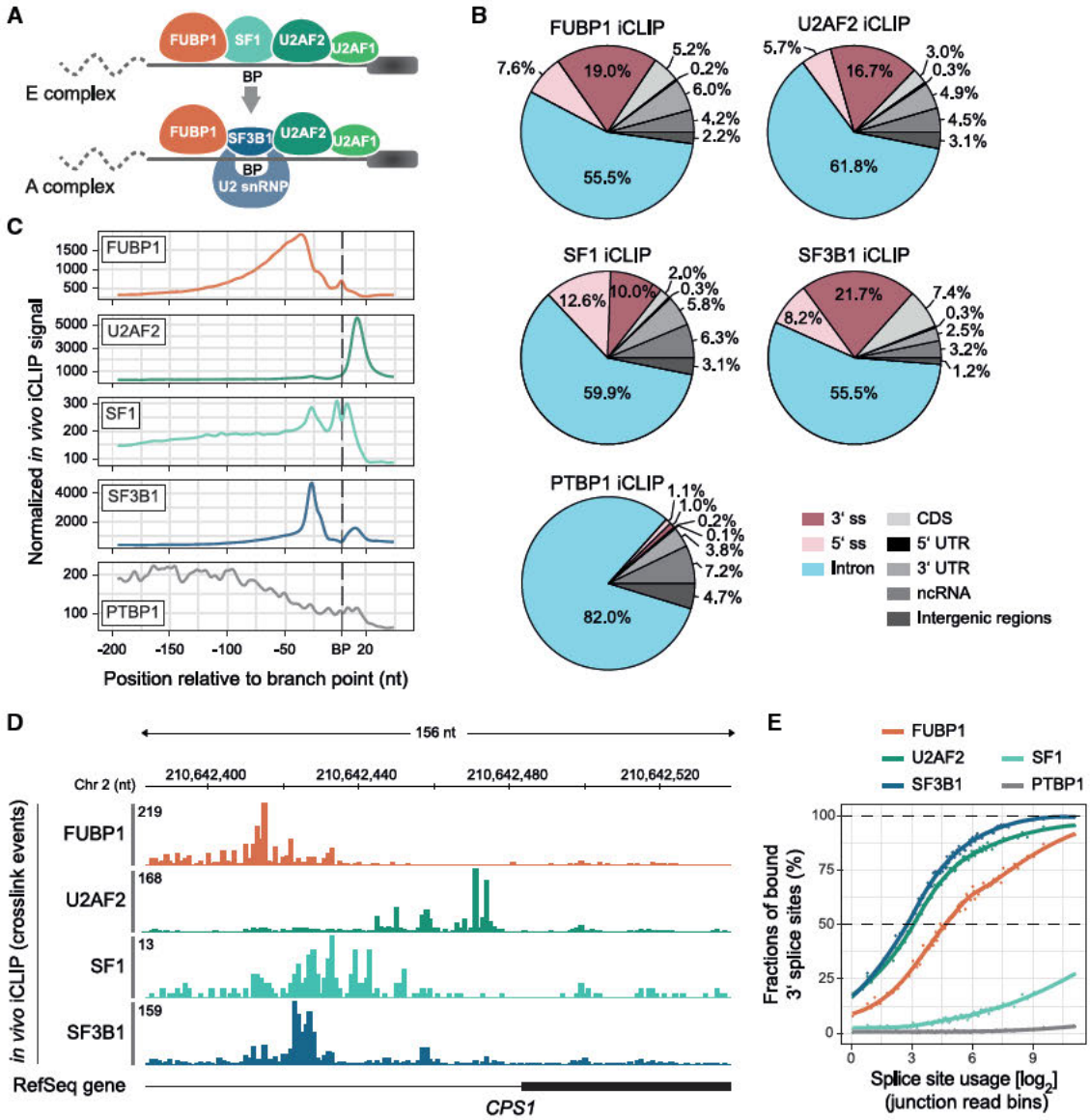


Figure 1. FUBP1 binds upstream of the branch point at 3′ splice sites during early spliceosome assembly *in vivo*

(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

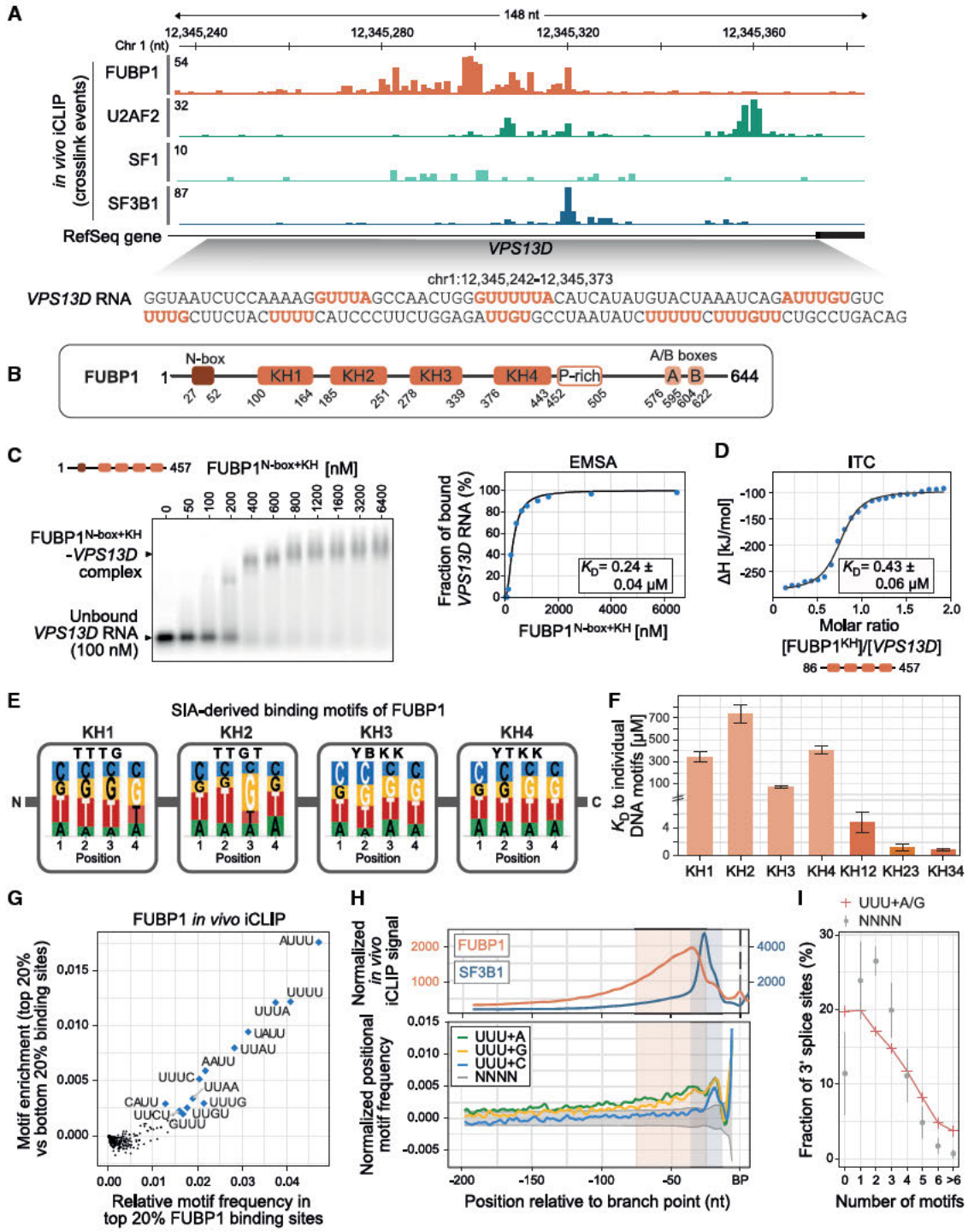


Figure 2. FUBP1 binds a hitherto unknown *cis*-regulatory motif upstream of the BP

(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).

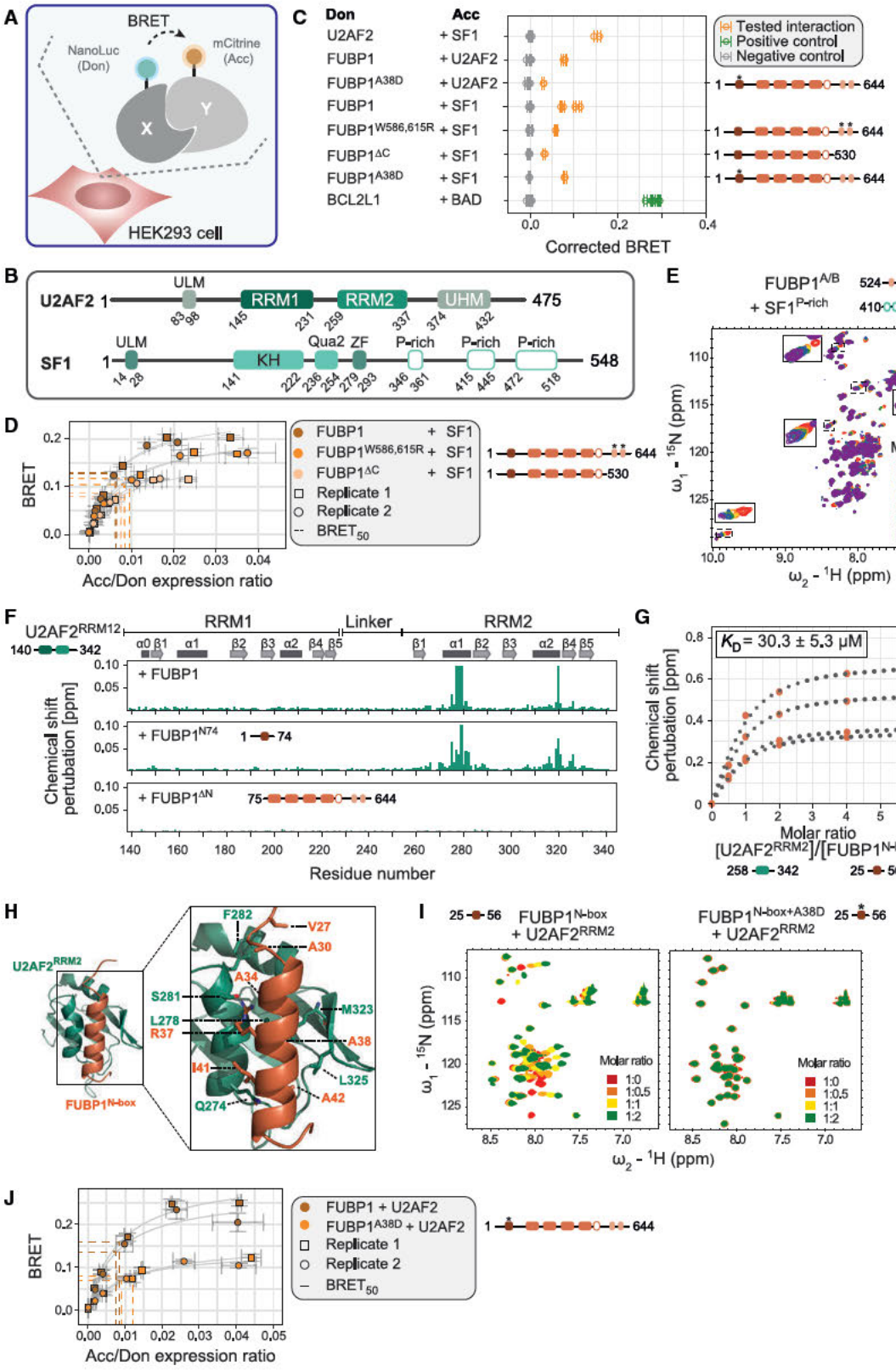
(E) Scaffold-independent analysis (SIA)-derived binding motifs for individual FUBP1 KH domains. Preferred bases are highlighted in white. Y, pyrimidine (T or C); B, not A (C, G, or T); K, keto (G or T).

(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).

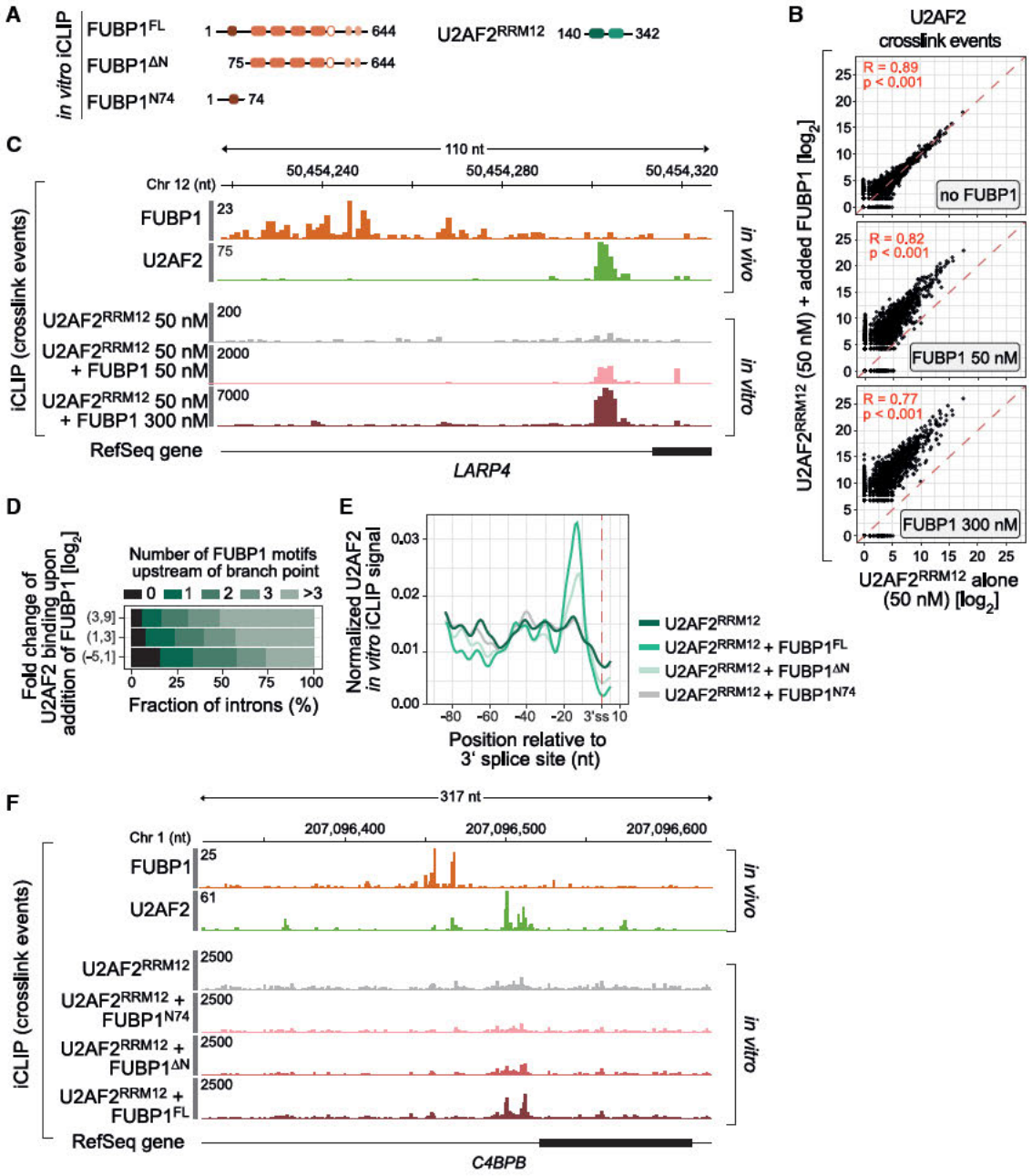


Figure 4. FUBP1 stabilizes U2AF2 binding at 3′ splice sites *in vitro*

(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$^{\Delta N}$.

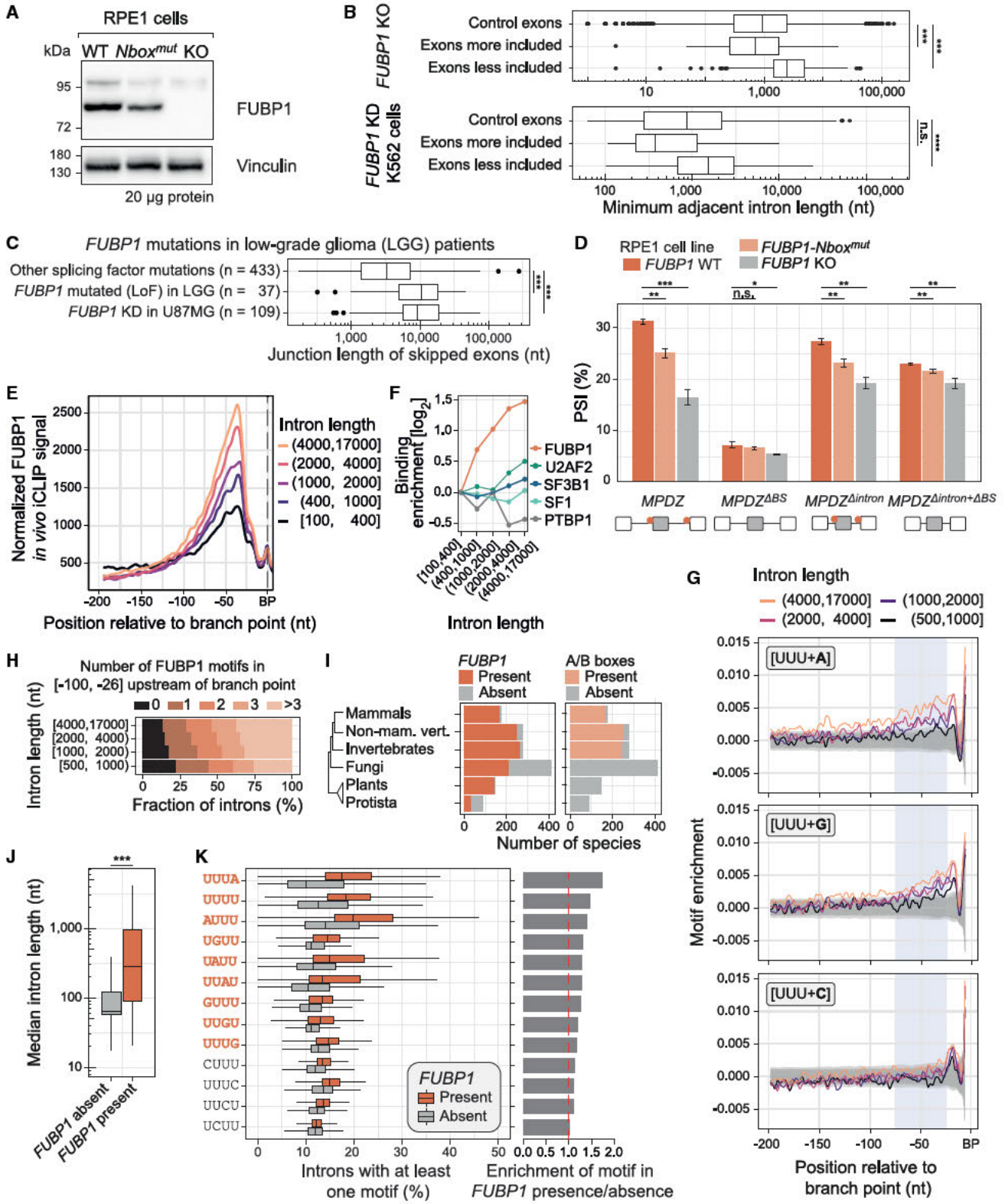


Figure 5. FUBP1 binds stronger to long introns and regulates exons flanked by long introns

(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).

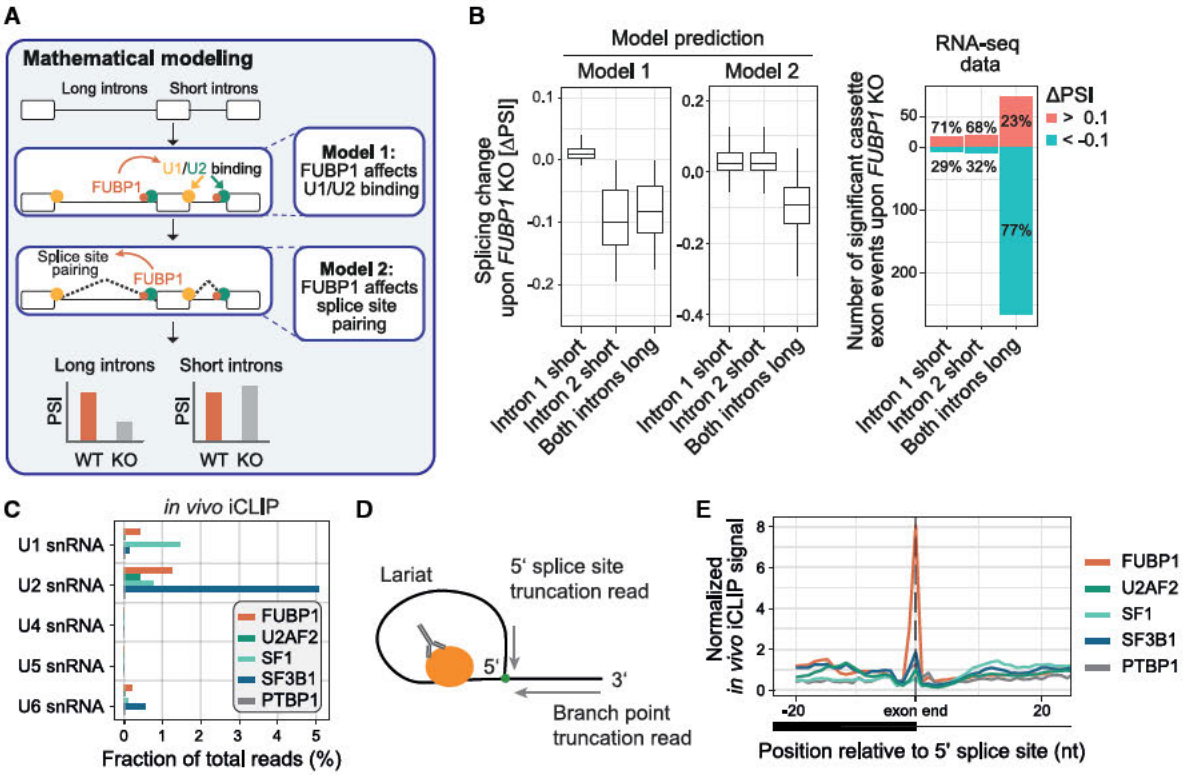
(G) Positional enrichment of FUBP1 binding motifs and control motifs relative to branch point and for various intron lengths. UUU+A/G/C, sets of four 4-mers containing UUU interspersed at any position with A/G/C. NNNN, 100,000 sets of random combinations of four 4-mers. 4-mer frequencies were calculated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

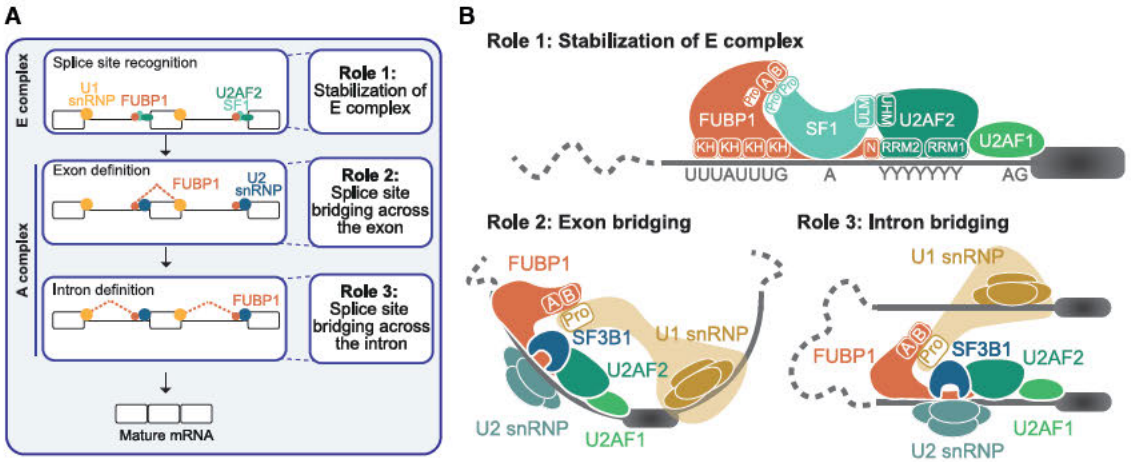


Figure 7. FUBP1 acts at multiple steps of early spliceosomal assembly

(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.

Molecular Cell, Volume *83*

Supplemental information

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

Stefanie Ebersberger, Clara Hipp, Miriam M. Mulorz, Andreas Buchbender, Dalmira Hubrich, Hyun-Seo Kang, Santiago Martínez-Lumbreras, Panajot Kristofori, F.X. Reymond Sutandy, Lidia Llacsahuanga Allcca, Jonas Schönfeld, Cem Bakisoglu, Anke Busch, Heike Hänel, Kerstin Tretow, Mareen Welzel, Antonella Di Liddo, Martin M. Möckel, Kathi Zarnack, Ingo Ebersberger, Stefan Legewie, Katja Luck, Michael Sattler, and Julian König

Figure S1

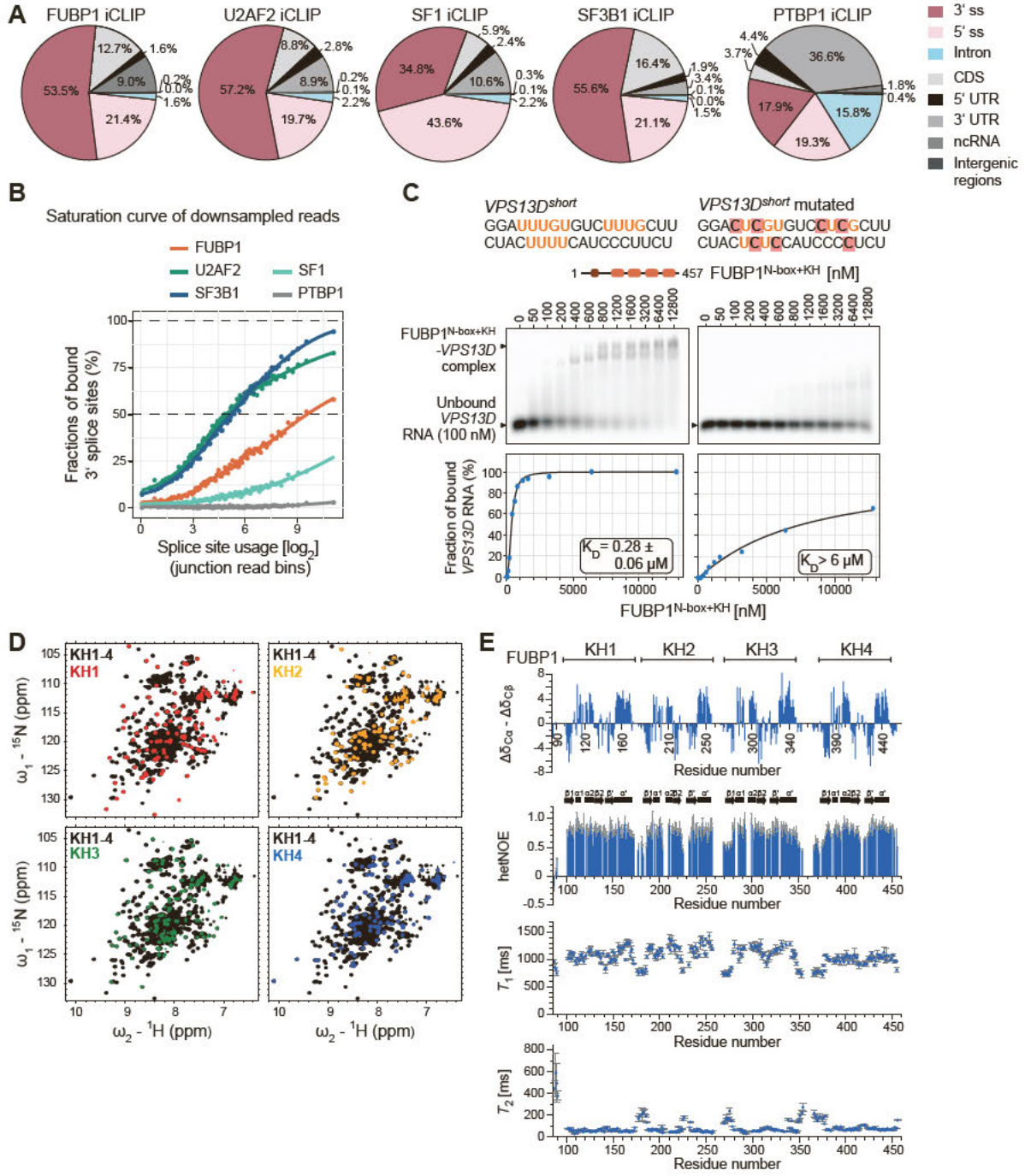


Figure S1. FUBP1 binding at 3' splice sites and RNA binding of KH domains (related to Figure 1B, 1E and 2B-D)

(A) 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) iCLIP in HeLa cells (normalized for total transcript length). 3' and 5' splice site (ss) refer to 100 nt upstream and downstream of exons, respectively. CDS, coding sequence; UTR, untranslated region.

(B) Saturation analysis on downsampled iCLIP data (FUBP1, ~57,000,000 crosslink events; SF3B1, ~68,000,000; U2AF2, 54,000,000; SF, 58,000,000; PTB, 49,000,000), where the iCLIP data for each splicing factor have approximately the same sequencing depth.

(C) Fluorescent electrophoretic mobility shift assay (EMSA) experiment on recombinant FUBP1$^{N\text{-}box+KH}$ (aa 1–457, 50 nM–12.8 μM) binding to a shortened 36-nt RNA fragment from *VPS13D* (*VPS13D*^{short}, 100 nM) (left). Agarose gel image (top) and quantification (bottom) with fitted curve show FUBP1–RNA binding in the nanomolar range (dissociation constant [K_D] = 0.28 ± 0.06 μM). Agarose gel of a fluorescent EMSA experiment on recombinant FUBP1$^{N\text{-}box+KH}$ (aa 1–457, 50 nM–12.8 μM) binding to *VPS13D*^{short} mutated (100 nM) with U-to-C mutations in U-rich stretches affording greatly reduced binding (right).

(D) Overlays of the ^{1}H–^{15}N heteronuclear single quantum coherence (HSQC) spectra of FUBP1 KH1–4 (black) with single KH domains (KH1, red; KH2, yellow; KH3, green; KH4, blue). Nuclear magnetic resonance (NMR) experiments of FUBP1^{KH} (KH1–4) show excellent spectral quality, despite the high molecular weight (~40 kDa), allowing most of the backbone chemical shifts to be assigned (310 out of 371 residues).

(E) $^{13}C_\alpha$ and $^{13}C_\beta$ secondary chemical shifts and ^{15}N relaxation experiments: $\{^{1}H\}$-^{15}N heteronuclear nuclear Overhauser effect (NOE), T_1, T_2, of the four KH domains of FUBP1. Folded KH domains exhibit more rigidity (NOE ~ 0.9, T_1 ~ 1s, T_2 ~ 60 ms) whereas linker regions are more flexible (lower NOE, lower T_1, higher T_2).

Figure S2

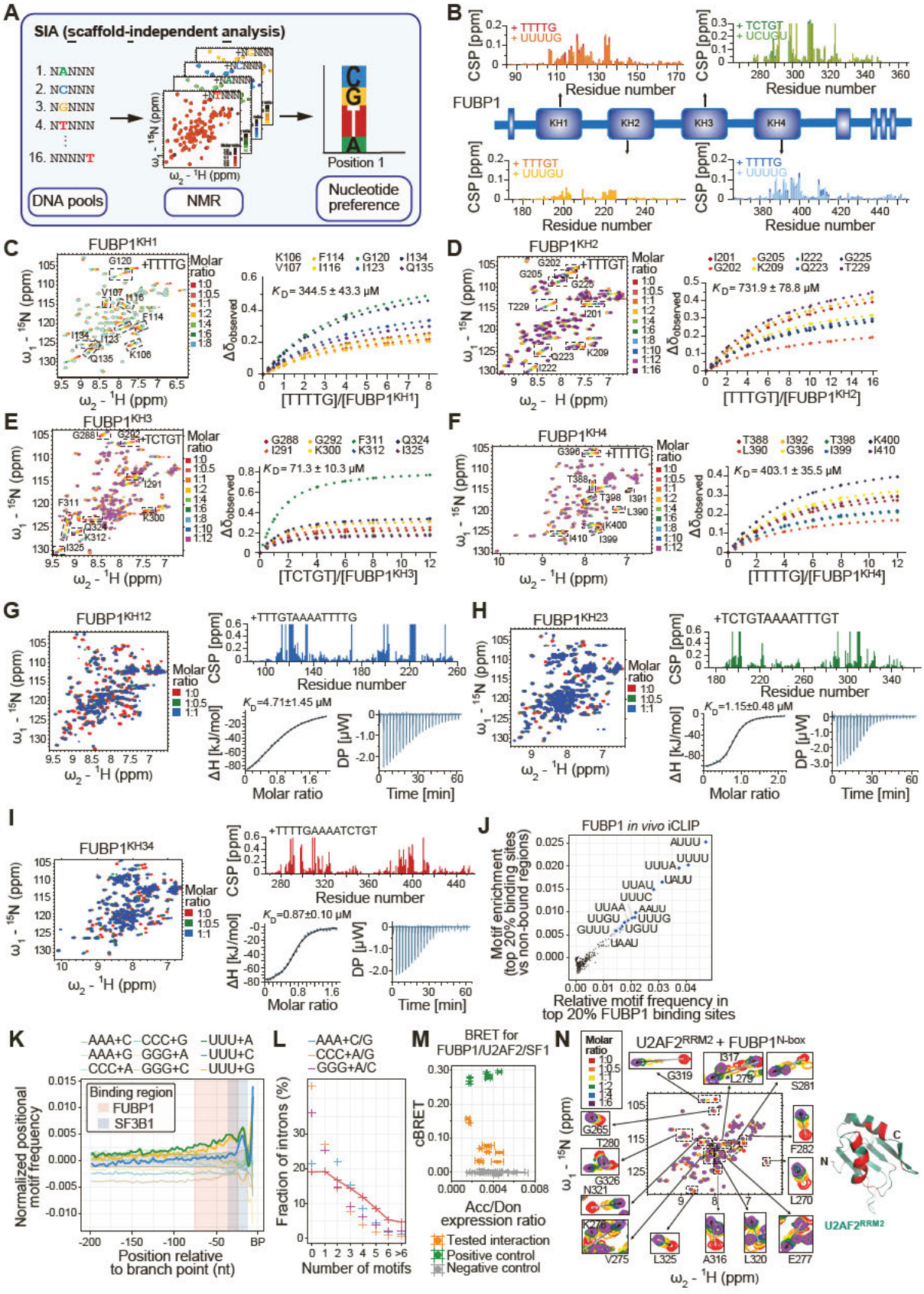

A
SIA (scaffold-independent analysis)
1. NANNN
2. NCNNN
3. NGNNN
4. NTNNN
16. NNNNT
DNA pools
NMR
Nucleotide preference
Position 1
B
FUBP1
KH1
KH2
KH3
KH4
+ TTTTG
+ UUUUG
+ TCTGT
+ UCUGU
+ TTTGT
+ UUUGU
CSP [ppm]
Residue number
C
FUBP1KH1
+TTTTG
Molar ratio
KD = 344.5 ± 43.3 µM
[TTTTG]/[FUBP1KH1]
D
FUBP1KH2
+TTTGT
KD = 731.9 ± 78.8 µM
[TTTGT]/[FUBP1KH2]
E
FUBP1KH3
+TCTGT
KD = 71.3 ± 10.3 µM
[TCTGT]/[FUBP1KH3]
F
FUBP1KH4
+TTTTG
KD = 403.1 ± 35.5 µM
[TTTTG]/[FUBP1KH4]
G
FUBP1KH12
+TTTGTAAAATTTTG
KD =4.71±1.45 µM
Molar ratio
Time [min]
H
FUBP1KH23
+TCTGTAAAATTTGT
KD =1.15±0.48 µM
I
FUBP1KH34
+TTTTGAAAATCTGT
KD =0.87±0.10 µM
J
FUBP1 in vivo iCLIP
Motif enrichment (top 20% binding sites vs non-bound regions)
Relative motif frequency in top 20% FUBP1 binding sites
K
Binding region
FUBP1
SF3B1
Normalized positional motif frequency
Position relative to branch point (nt)
L
Fraction of introns (%)
Number of motifs
M
BRET for FUBP1/U2AF2/SF1
cBRET
Acc/Don expression ratio
Tested interaction
Positive control
Negative control
N
U2AF2RRM2 + FUBP1N-box
U2AF2RRM2

Figure S2. Scaffold-independent analysis and titration curves for the final optimal binding motifs for each KH domain (related to Figure 2E-I, 3C, F)

(A) Schematic workflow of the NMR-based scaffold-independent analysis (SIA) [S1]. SIA reports on the nucleic acid binding specificity of a given RNA-binding protein (RBP) at each position of a nucleic acid target. Sixteen 5-mer DNA pools with one specific nucleotide fixed at one position, otherwise randomized, are individually titrated to each KH domain. The observed changes in chemical shift of the selected peaks are averaged and normalized for each DNA pool to obtain a score for the nucleotide position and type preference.

(B) Comparisons of chemical shift perturbations (CSPs) of all four FUBP1 KH domains upon addition of the optimal nucleotide motifs as either DNA or RNA (1:1 molar ratio of protein to RNA). This shows that DNA and RNA binding are very similar for all four KH domains of FUBP1.

(C) NMR titration and dissociation constant (K_D) calculation for the binding of FUBP1 KH1 (100 μM) with TTTTG up to a protein/DNA molar ratio of 1:8. The indicated residues are used for the calculation of K_D. As expected for interactions of KH domains to nucleic acids, the changes in chemical shift are mostly mapped to the α1 and α2 helices and the GXXG loop.

(D) NMR titration and K_D calculation for the binding of FUBP1 KH2 (100 μM) with TTTGT up to a protein/DNA molar ratio of 1:16. The marked residues are used for the calculation of K_D. As expected, the changes in chemical shift are mostly mapped to the α1 and α2 helices and the GXXG loop.

(E) NMR titration and K_D calculation for the binding of FUBP1 KH3 (100 μM) with TCTGT up to a protein/DNA molar ratio of 1:12. The marked residues are used for the calculation of K_D. As expected, the changes in chemical shift are mostly mapped to the α1 and α2 helices and the GXXG loop.

(F) NMR titration and K_D calculation for the binding of FUBP1 KH4 (100 μM) with TTTTG up to a protein/DNA molar ratio of 1:12. The marked residues are used for the calculation of K_D. As expected, the changes in chemical shift are mostly mapped to the α1 and α2 helices and the GXXG loop.

(G) NMR titration and ITC of FUBP1 KH1–2 binding to a DNA oligonucleotide containing an optimal DNA motif for each KH domain derived by SIA linked by an A4 linker (TTTGTAAAATTTTG). Consistent with the K_D values from ITC, the NMR titration indicates binding in an intermediate exchange regime.

(H) NMR titration and ITC of FUBP1 KH2-3 binding to a DNA oligonucleotide containing an optimal DNA motif for each KH domain derived by SIA linked by an A4 linker (TCTGTAAAATTTGT). Consistent with the K_D values from ITC, the NMR titration indicates binding in an intermediate exchange regime.

(I) NMR titration and ITC of FUBP1 KH3–4 binding to a DNA oligonucleotide containing an optimal DNA motif for each KH domain derived by SIA linked by an A4 linker (TTTTGAAAATCTGT). Consistent with the K_D values from ITC, the NMR titrations indicate binding in an intermediate exchange regime.

(J) Motif enrichment in the *in vivo* FUBP1 iCLIP data. Disjunct 4-mer frequencies were calculated in extended windows (5-nt binding site ± 5 nt) around the top 20% of binding sites based on expression-normalized iCLIP signal and in non-bound regions in the same introns excluding a 20-nt region downstream of the 5' ss and a 150-nt region upstream of the branch point (BP). Enrichment for each motif is defined as the distance for each data point to the diagonal in the scatterplot of relative motif frequencies of the top 20% vs bottom 20% of binding sites.

(K) Positional enrichment of FUBP1 binding motifs and control motifs relative to the BP. UUU+A/G/C, that is, 4-mers containing UUU interspersed at any position with A/G/C. Control motif sets are mononucleotide tracts interspersed by one other nucleotide. 4-mer frequencies were calculated position-wise upstream of the BP and compared to the average 4-mer frequencies in an

intronic control region (a 100-nt-long region 100 nt downstream of the 5' splice site). Shaded regions correspond to the main binding regions of FUBP1 (red) and SF3B1 (blue).

(L) Abundance of FUBP1 binding motifs (UUU+A/G) at 3' ss of human introns. Abundance for other mononucleotide motifs (AAA + C/G & AAAA, CCC + G/C & CCCC, GGG + A/C & GGGG) is given for the purpose of comparison.

(M) Total luminescence and fluorescence measurements were used to estimate the amounts of FUBP1 or the mutants $FUBP1^{A38D}$, $FUBP1^{\Delta C}$, and $FUBP1^{W586,615R}$ paired with wild-type U2AF2 and SF1 (orange), BCL2L1-BAD as a positive control pair (green) and pairs that are not known to interact with each other as negative controls (gray) in bioluminescence resonance energy transfer (BRET)-based assay. Acceptor/donor ratios are similar for all pairs, making the cBRET values more comparable to each other.

(N) NMR titration of $U2AF2^{RRM2}$ with $FUBP1^{N\text{-}box}$ up to sixfold molar excess (left). Significantly shifted peaks are enlarged. The peaks with a chemical shift perturbation (CSP) ≥ 0.1 are shown in red along with corresponding residues on the structure of $U2AF2^{RRM2}$ (right) (PDB ID: 8P25).

Figure S3

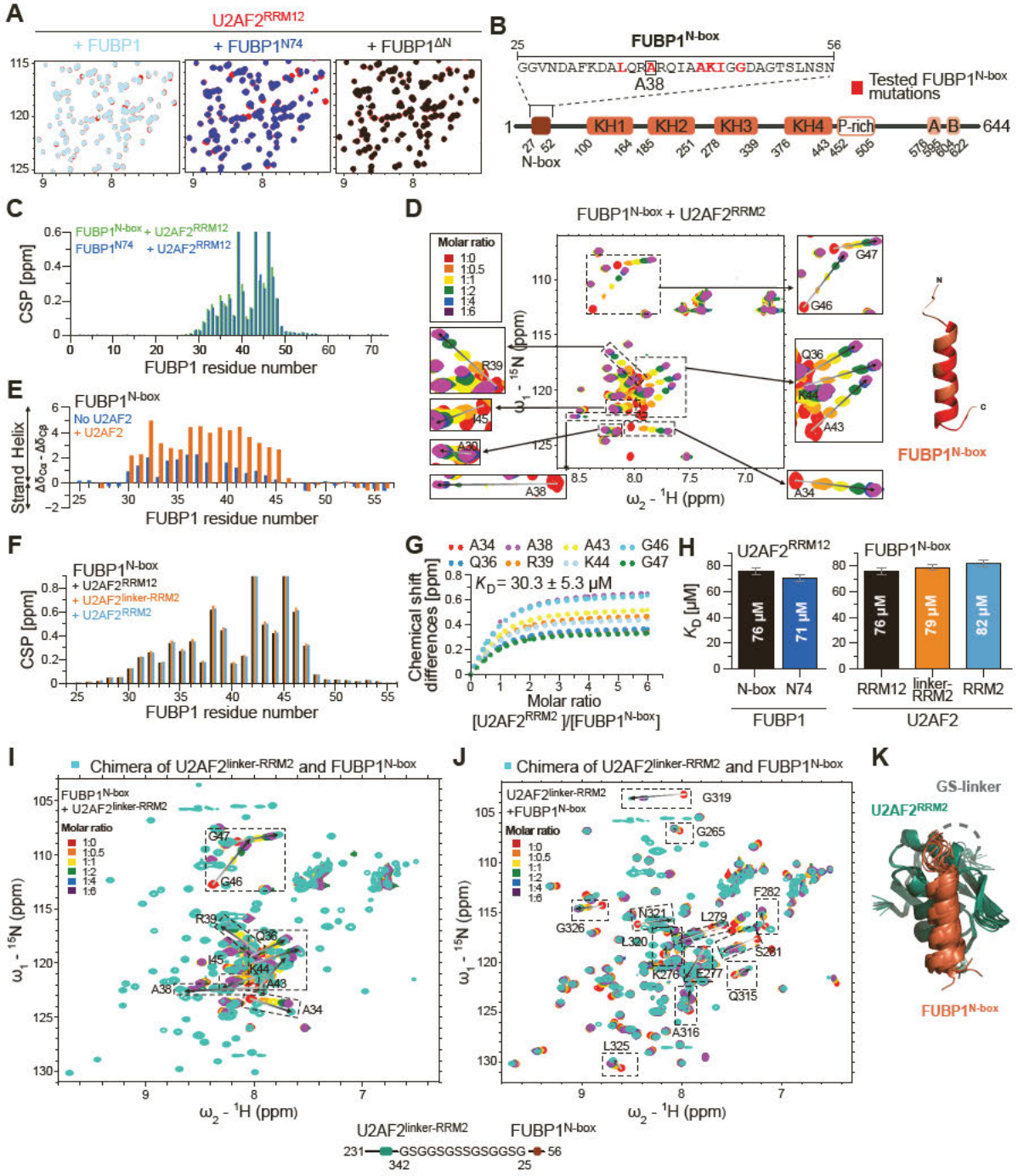


Figure S3. Determination of the minimal interaction interface between FUBP1$^{N\text{-box}}$ and U2AF2^{RRM2} (related to Figure 3F-H)

(A) Comparison of a selected region of the one-point NMR titrations (0.5 molar ratio) of U2AF2^{RRM2} (red) with full-length FUBP1 (cyan), FUBP1^{N74} (blue), and FUBP1$^{\Delta N}$ (black) showing significant chemical shift changes.

(B) Overview of the FUBP1 construct used for NMR and BRET experiments. Red color marks mutations that are tested for effect on binding of FUBP1 with U2AF2.

(C) Comparison of CSPs upon the titration of a shortened FUBP1 N-terminal construct (FUBP1$^{N\text{-box}}$, aa 25–56; green) and FUBP1^{N74} (blue) with U2AF2^{RRM12}.

(D) NMR titration of FUBP1$^{N\text{-box}}$ with U2AF2^{RRM2} up to sixfold molar excess (left). Significantly shifted peaks are boxed and enlarged. The peaks with a CSP ≥ 0.1 are highlighted in red along with corresponding residues on the structure of FUBP1$^{N\text{-box}}$ (right) (PDB ID: 8P25).

(E) Comparison of the C_α and C_β chemical shift-derived secondary structure of free FUBP1$^{N\text{-box}}$ (blue) and FUBP1$^{N\text{-box}}$ bound to U2AF2^{RRM2} (orange). The fractional helical conformation for residues 30–45 in the absence of U2AF2 is further increased upon binding to U2AF2.

(F) Comparison of the CSP of FUBP1$^{N\text{-box}}$ titrations with U2AF2 constructs of various lengths (U2AF2^{RRM12}, black; U2AF2$^{linker\text{-}RRM2}$, orange; U2AF2^{RRM2}, light blue).

(G) Calculation of K_D for the FUBP1$^{N\text{-box}}$ and U2AF2^{RRM2} interaction derived by NMR titration. The changes in chemical shift of selected residues in the titration of FUBP1$^{N\text{-box}}$ with U2AF2^{RRM2} (shown in panel D) are plotted against the molar ratio of ligand to titrant.

(H) Comparison of K_D values for the interaction of FUBP1$^{N\text{-box}}$ (black) and FUBP1^{N74} (blue) with U2AF2^{RRM12} and FUBP1$^{N\text{-box}}$ with U2AF2^{RRM12} (black), U2AF2$^{linker\text{-}RRM2}$ (orange), and U2AF2^{RRM2} (light blue), determined by ITC (**Table S2**). The measurements were performed in triplicates and data are represented as mean ± SD.

(I) Overlay of the ^{1}H–^{15}N HSQC spectra of the chimeric construct U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-box}}$ (cyan) and FUBP1$^{N\text{-box}}$ titrated with U2AF2$^{linker\text{-}RRM2}$ (molar ratio of 1:6).

(J) Overlay of the ^{1}H–^{15}N HSQC spectra of the chimeric construct U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-box}}$ (cyan) and U2AF2$^{linker\text{-}RRM2}$ titrated with FUBP1$^{N\text{-box}}$ (molar ratio of 1:6).

(K) NMR ensemble (10 lowest energy structures) of the chimeric construct U2AF2$^{linker\text{-}RRM2}$ (green)/FUBP1$^{N\text{-box}}$ (brown). The end of the flexible linker between RRM1-RRM2 (231–245) is not shown, the artificial GS-linker between the C terminus of U2AF2 RRM2 and the N-terminal region of FUBP1$^{N\text{-box}}$ are indicated by gray dashed lines (PDB ID: 8P25).

Figure S4

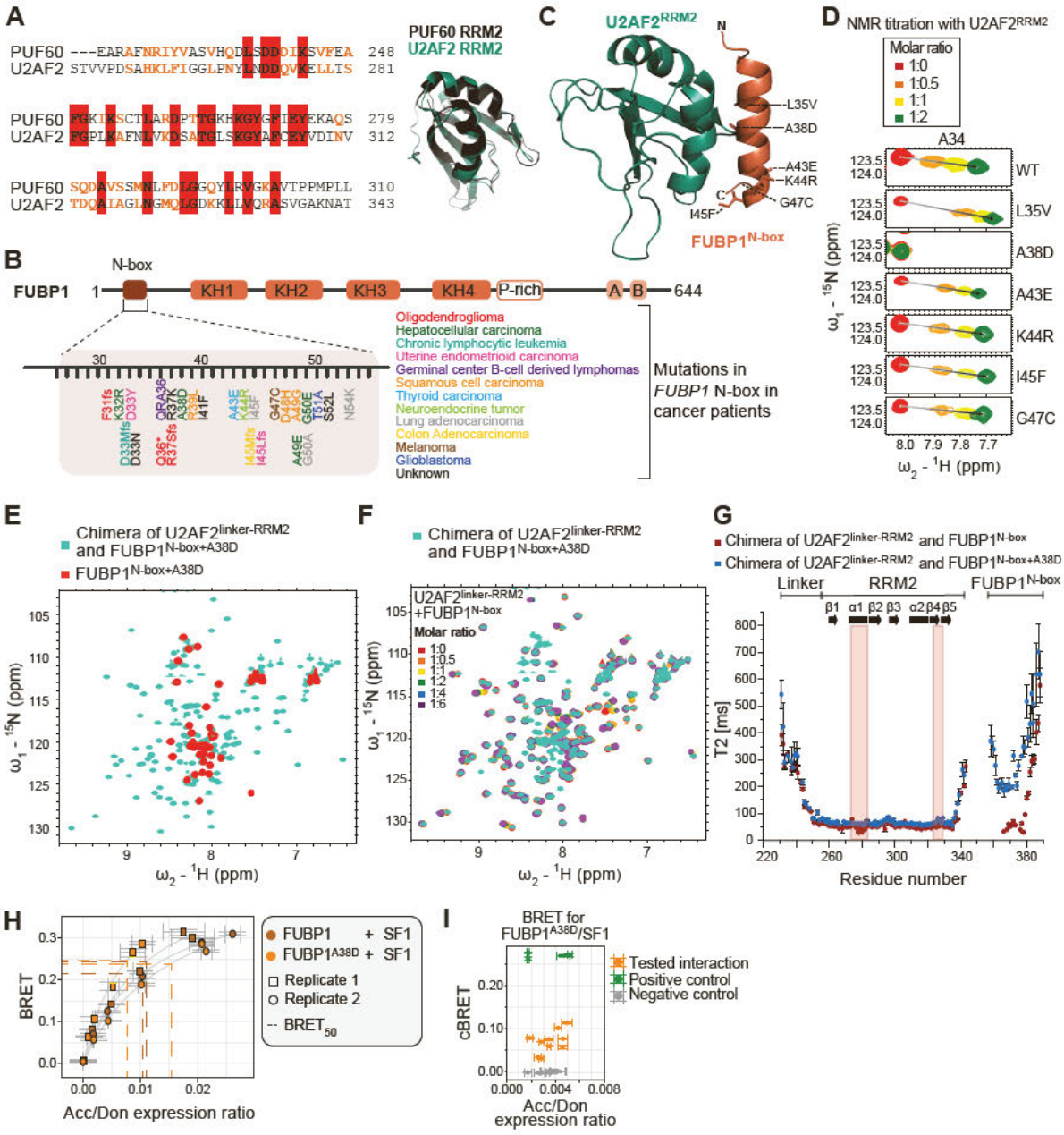


Figure S4. The effects of cancer-related mutations in the FUBP1 N-box on the interaction with U2AF2 RRM2 (related to Figure 3C, 3I-J)

(A) Sequential alignment (Clustal Omega [S4]) of the RRM2 domains of human PUF60 and U2AF2, mapping conserved residues (red) and similar residues (orange). Overlay of the structure of PUF60 RRM2 (black) and U2AF2 RRM2 (green) (adapted from [S2] and [S3]; PDB IDs: 2KXH, 6TR0).

(B) FUBP1 N-box mutations identified in different cancer types. Databases (see STAR Methods) were screened for the occurrence of cancer-related mutations within the region of *FUBP1* encoding for the N-box, yielding one insertion, five frameshifts (fs) leading to a premature termination codon (*) and 20 missense variants.

(C) Cancer-related mutations (labeled and side chains shown on the calculated structure of a chimeric construct of U2AF2^{RRM2} and FUBP1$^{N\text{-}box}$, PDB ID: 8P25) within the helical binding region of FUBP1$^{N\text{-}box}$ and located at the interfaces with U2AF2^{RRM2} were selected for further NMR study.

(D) Comparison of the changes in chemical shift of residue A34 for the titration of FUBP1$^{N\text{-}box}$ wild-type and mutants (L35V, A38D, A43E, K44R, I45F, G47C) upon adding U2AF2^{RRM2}.

(E) Overlay of the ^{1}H–^{15}N HSQC spectra of the A38D mutant of FUBP1$^{N\text{-}box}$ (red) with the chimeric construct U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-}box}$ with A38D mutation (cyan).

(F) Overlay of the ^{1}H–^{15}N HSQC spectra of the chimeric construct U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-}box}$ with A38D mutation (cyan) with the titration of FUBP1$^{N\text{-}box}$ with U2AF2$^{linker\text{-}RRM2}$ shows that the mutant spectrum resembles those of the unbound individual components.

(G) Comparison of the ^{15}N T_2 relaxation rates of the chimeric constructs U2AF2$^{linker\text{-}RRM2}$/FUBP1$^{N\text{-}box}$ wild-type and A38D mutant. Increased T_2 relaxation rates in the N-box helix of the A38D mutant chimera compared to the wild-type is consistent with much weaker binding of the mutant to the U2AF2 RRM2.

(H) BRET titration curves shown for FUBP1 and FUBP1^{A38D} versus SF1. As expected, mutation of the FUBP1 N-box does not result in significant loss of binding to SF1. Two biological replicates are shown, each done in technical triplicates. Error bars represent the standard deviation.

(I) Total luminescence (Don) and fluorescence (Acc) ratios were determined for FUBP1 and FUBP1^{A38D} versus SF1. Acceptor/donor ratios are similar for all pairs making the cBRET values more comparable to each other.

Figure S5

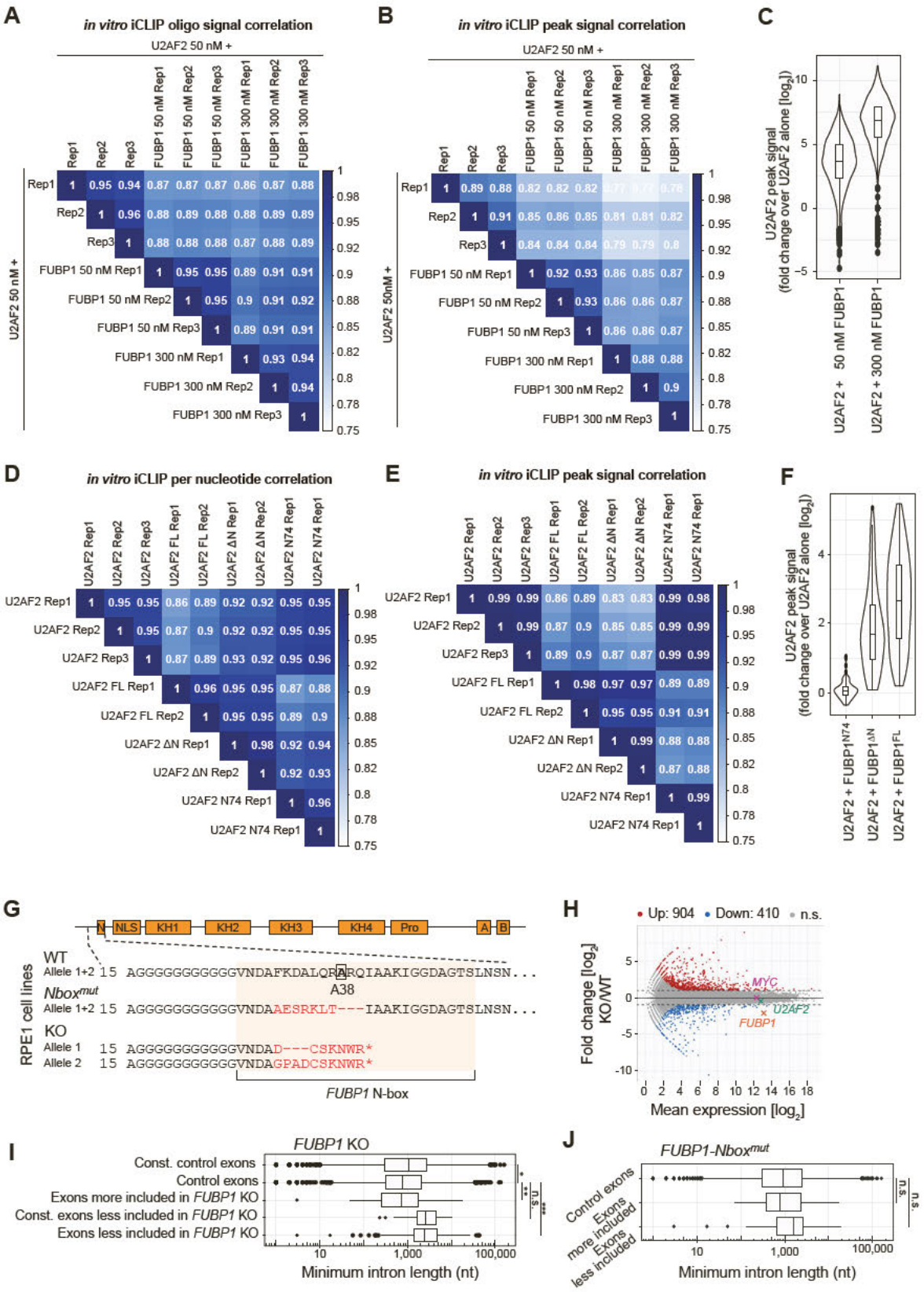

A
in vitro iCLIP oligo signal correlation
U2AF2 50 nM +
Rep1
Rep2
Rep3
FUBP1 50 nM Rep1
FUBP1 50 nM Rep2
FUBP1 50 nM Rep3
FUBP1 300 nM Rep1
FUBP1 300 nM Rep2
FUBP1 300 nM Rep3
U2AF2 50 nM +
B
in vitro iCLIP peak signal correlation
U2AF2 50 nM +
U2AF2 50nM +
C
U2AF2 peak signal (fold change over U2AF2 alone [log2])
U2AF2 + 50 nM FUBP1
U2AF2 + 300 nM FUBP1
D
in vitro iCLIP per nucleotide correlation
U2AF2 Rep1
U2AF2 Rep2
U2AF2 Rep3
U2AF2 FL Rep1
U2AF2 FL Rep2
U2AF2 ΔN Rep1
U2AF2 ΔN Rep2
U2AF2 N74 Rep1
U2AF2 N74 Rep1
E
in vitro iCLIP peak signal correlation
F
U2AF2 peak signal (fold change over U2AF2 alone [log2])
U2AF2 + FUBP1N74
U2AF2 + FUBP1ΔN
U2AF2 + FUBP1FL
G
NLS
KH1
KH2
KH3
KH4
Pro
A
B
RPE1 cell lines
WT
Allele 1+2 15 AGGGGGGGGGGGVNDAFKDALQRARQIAAKIGGDAGTSLNSN...
A38
Nbox^mut
Allele 1+2 15 AGGGGGGGGGGGVNDAAESRKLT---IAAKIGGDAGTSLNSN...
KO
Allele 1 15 AGGGGGGGGGGGVNDAD---CSKNWR*
Allele 2 15 AGGGGGGGGGGGVNDAGPADCSKNWR*
FUBP1 N-box
H
Up: 904
Down: 410
n.s.
Fold change [log2] KO/WT
Mean expression [log2]
MYC
U2AF2
FUBP1
I
FUBP1 KO
Const. control exons
Control exons
Exons more included in FUBP1 KO
Const. exons less included in FUBP1 KO
Exons less included in FUBP1 KO
Minimum intron length (nt)
J
FUBP1-Nbox^mut
Control exons
Exons more included
Exons less included
Minimum intron length (nt)

Figure S5. Reproducibility between replicates and changes in U2AF2^{RRM12} binding from *in vitro* iCLIP experiments and expression and splicing changes upon *FUBP1* KO (related to Figure 4A-F, 5A-B)

(A) Reproducibility of *in vitro* iCLIP data with oligonucleotide-derived transcript library. The correlation matrix shows pairwise Pearson correlation of U2AF2^{RRM12} crosslink events per oligonucleotide (n = 1,998) between samples. Experiments were performed with U2AF2^{RRM12} alone (50 nM) and with the addition of full-length FUBP1 at 50 or 300 nM.

(B) Reproducibility of *in vitro* iCLIP data with oligonucleotide-derived transcript library. The correlation matrix shows pairwise Pearson correlation of total U2AF2^{RRM12} crosslink events inside U2AF2 binding sites between samples (1,831 oligonucleotides harbor a U2AF2 binding sites according to U2AF2 *in vivo* iCLIP). Experiments as in panel A.

(C) Comparative boxplot of normalized U2AF2^{RRM12} crosslink events per binding site between conditions (n = 1,504). Experiments as in panel A.

(D) Reproducibility of *in vitro* iCLIP data with eight long *in vitro* transcripts [S5]. The correlation matrix shows pairwise Pearson correlation of U2AF2^{RRM12} crosslink events per nucleotide over all *in vitro* transcripts between samples. Experiments were performed with U2AF2^{RRM12} alone (50 nM) and with the addition of full-length FUBP1^{FL}, FUBP1^{N74}, and FUBP1$^{\Delta N}$ (all 50 nM).

(E) Reproducibility of *in vitro* iCLIP data with eight long *in vitro* transcripts [S5]. Correlation matrix shows pairwise Pearson correlation of total binding signals (n = 109) between samples. Experiments as in panel D.

(F) Comparative boxplot of normalized U2AF2^{RRM12} crosslink events between conditions (n = 109). Experiments as in panel D.

(G) Zoom-in of the FUBP1$^{N\text{-}box}$ sequence, which when targeted with CRISPR/Cas9 results in a knockout cell line (*FUBP1* KO) and a mutant cell line (*FUBP1-Nbox^{mut}*), in which FUBP1 lacks the U2AF2 interaction surface.

(H) Log_2 fold change versus mean expression for genes upon *FUBP1* KO in RPE1 cells.

(I) Minimum adjacent intron length for cassette exons that are more or less included and for constitutive exons less included in *FUBP1-Nbox^{mut}* RPE1 cells (n = 123/249/27) compared to unchanged control exons (n = 4,584) and unchanged constitutive control exons (n = 5,717).

(J) Minimum adjacent intron length for cassette exons that are more or less included in *FUBP1-Nbox^{mut}* RPE1 cells (n = 36/45) compared to unchanged control exons (n = 10,678).

Figure S6

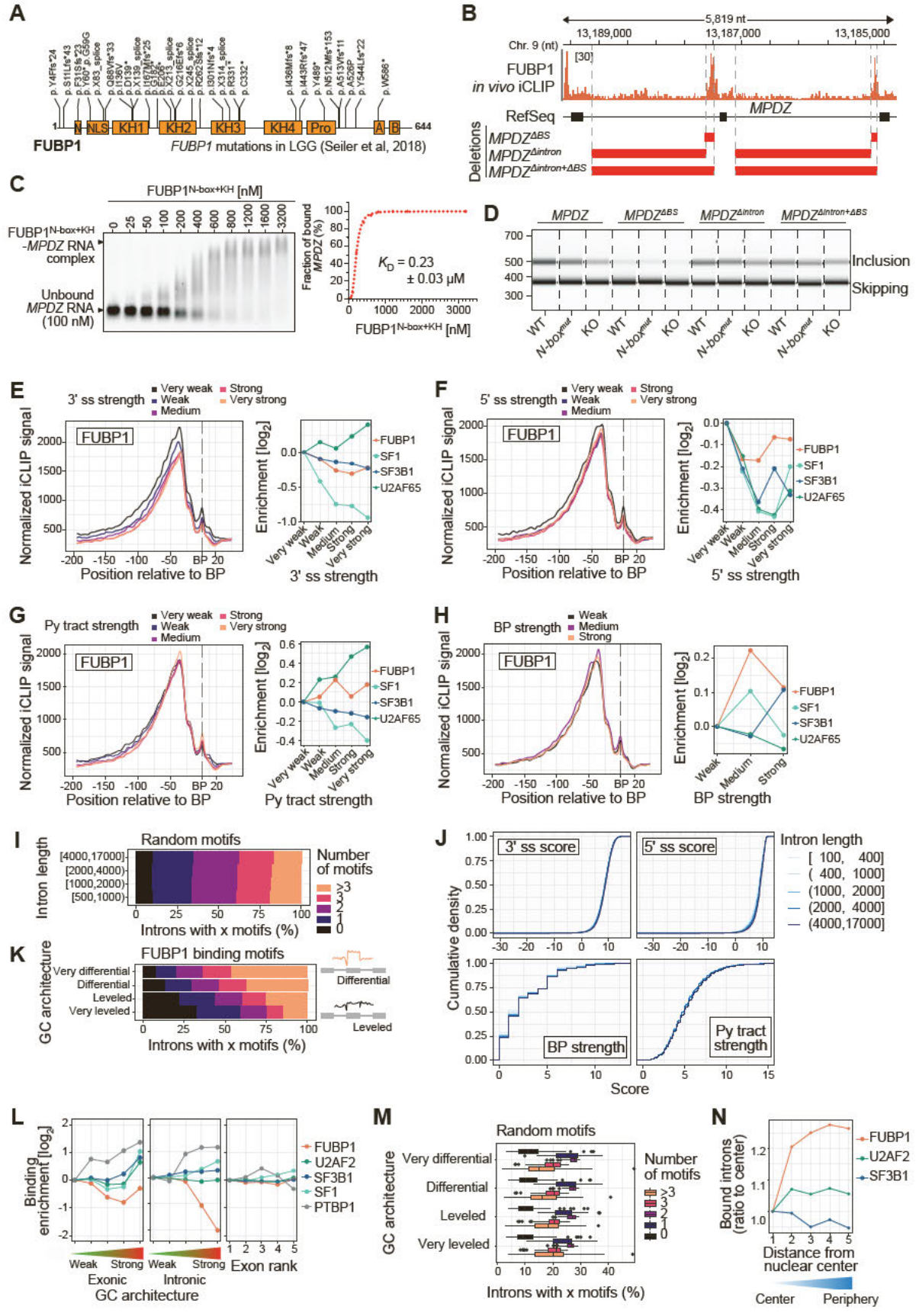

A
FUBP1
FUBP1 mutations in LGG (Seiler et al, 2018)
B
5,819 nt
13,189,000
13,187,000
13,185,000
Chr. 9 (nt)
FUBP1
in vivo iCLIP
RefSeq
MPDZ
Deletions
C
FUBP1N-box+KH [nM]
FUBP1N-box+KH -MPDZ RNA complex
Unbound MPDZ RNA (100 nM)
Fraction of bound MPDZ (%)
KD = 0.23 ± 0.03 μM
FUBP1N-box+KH [nM]
D
Inclusion
Skipping
WT
N-boxmut
KO
E
3' ss strength
Very weak
Weak
Medium
Strong
Very strong
FUBP1
Normalized iCLIP signal
Position relative to BP
Enrichment [log2]
SF1
SF3B1
U2AF65
F
5' ss strength
G
Py tract strength
H
BP strength
I
Random motifs
Intron length
Number of motifs
Introns with x motifs (%)
J
3' ss score
5' ss score
Intron length
Cumulative density
BP strength
Py tract strength
Score
K
FUBP1 binding motifs
GC architecture
Very differential
Differential
Leveled
Very leveled
L
Binding enrichment [log2]
U2AF2
PTBP1
Exonic
Intronic
Exon rank
M
N
Bound introns (ratio to center)
Distance from nuclear center
Center
Periphery

Figure S6. FUBP1 effects on long introns (related to Figure 5C-H)

(A) Position and identity of FUBP1 loss-of-function (LoF) mutations in glioma patients with 1p/19q deletion-positive background [S6].

(B) Genome browser view of the region included in the *MPDZ* minigene displaying the *in vivo* iCLIP data (crosslink events per nucleotide) of FUBP1 (orange). Deletions of introns with/without FUBP1 binding sites are indicated below with red bars.

(C) EMSA experiment to demonstrate binding of recombinant $FUBP1^{N\text{-}box+KH}$ (aa 1–457, 25–3200 nM) to a fluorescently labeled 132-nt RNA fragment from *MPDZ* (100 nM). Agarose gel image (bottom) and quantification (top) with fitted curve show FUBP1–RNA binding in a nanomolar range (K_D = 0.23 ± 0.03 μM).

(D) Capillary electrophoresis of exon inclusion levels upon intron shortening in the *MPDZ* minigene.

(E) Metaprofile showing the number of crosslink events of FUBP1 relative to the branch point in dependency on 3' splice site strength. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns (left). Binding enrichment quantification: Area under the curve (AUC) in each intron class compared to the AUC in introns with very low 3' splice site strength (right).

(F) Metaprofile showing the number of crosslink events of FUBP1 relative to the branch point in dependency on 5' splice site strength. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns (left). Binding enrichment quantification: AUC in each intron class compared to the AUC in introns with very low 5' splice site strength (right).

(G) Metaprofile showing the number of crosslink events of FUBP1 relative to the branch point in dependency on Py tract strength. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns (left). Binding enrichment quantification: AUC in each intron class compared to the AUC in introns with very low Py tract strength (right).

(H) Metaprofile showing the number of crosslink events of FUBP1 relative to the branch point in dependency on BP strength. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns (left). Binding enrichment quantification: AUC in each intron class compared to the AUC in introns with very weak BP strength (right).

(I) Fraction of introns with 0, 1, 2, 3, or > 3 motif sets of size 9 of random 4-mers in dependency on intron length. Random sets were drawn 100 times and the resulting fractions were then averaged.

(J) Cumulative distribution of splice site features conditioned on intron length.

(K) Number of FUBP1-binding motifs upstream of the BP ([−100 nt; −26 nt]) in dependency on differential GC content. Differential GC content is the GC content of the exon minus that of the first 100 nt of the downstream intron.

(L) Enrichment of FUBP1 binding upstream of the branch point in dependency on exon/intron GC content and exon rank. In the underlying metaprofiles, iCLIP signals are normalized for expression and then averaged per nucleotide over all introns.

(M) Fraction of introns with 0, 1, 2, 3 or > 3 motif sets of size 9 of random 4-mers in dependency on differential GC content. Random sets are drawn 100 times and resulting fractions are then averaged.

(N) Percent of introns bounds through different scopes of Euclidean distances where 1 means the nuclear center and 5 is the periphery. Enrichment is shown compared to the first scope. Based on data from [S7].

Figure S7

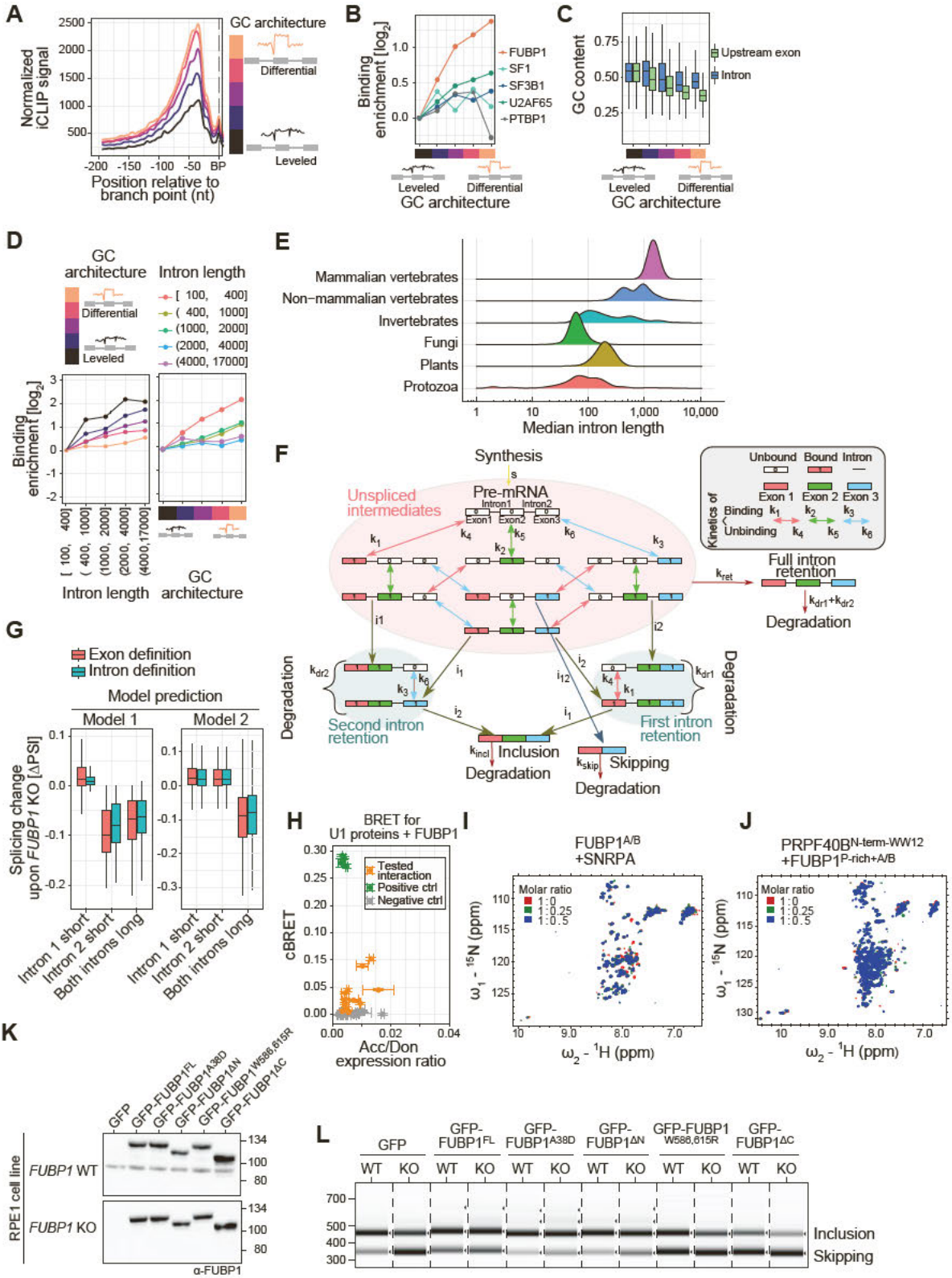

A
Normalized iCLIP signal
Position relative to branch point (nt)
GC architecture
Differential
Leveled
B
Binding enrichment [log2]
FUBP1
SF1
SF3B1
U2AF65
PTBP1
GC architecture
C
GC content
Upstream exon
Intron
D
Intron length
E
Mammalian vertebrates
Non−mammalian vertebrates
Invertebrates
Fungi
Plants
Protozoa
Median intron length
F
Synthesis
Pre-mRNA
Unspliced intermediates
Full intron retention
Degradation
Second intron retention
First intron retention
Inclusion
Skipping
Kinetics of
Unbound
Bound
Intron
G
Exon definition
Intron definition
Model prediction
Model 1
Model 2
Splicing change upon FUBP1 KO [ΔPSI]
Intron 1 short
Intron 2 short
Both introns long
H
BRET for U1 proteins + FUBP1
Tested interaction
Positive ctrl
Negative ctrl
cBRET
Acc/Don expression ratio
I
FUBP1A/B +SNRPA
Molar ratio
J
PRPF40BN-term-WW12 +FUBP1P-rich+A/B
K
GFP
GFP-FUBP1FL
GFP-FUBP1A38D
GFP-FUBP1ΔN
GFP-FUBP1W586,615R
GFP-FUBP1ΔC
RPE1 cell line
FUBP1 WT
FUBP1 KO
α-FUBP1
L
WT
KO
Inclusion
Skipping

Figure S7. Characterization and modeling of FUBP1 binding behavior (related to Figure 5E, 5I-J, 6A-B, 6G-I)

(A) Metaprofile showing the number of crosslink events of FUBP1 relative to the BP in dependency on differential GC content. iCLIP signals are normalized for expression and then averaged per nucleotide over all introns.

(B) Binding enrichment quantification: AUC in each intron class compared to the AUC in introns with leveled GC content.

(C) Comparison of exon and intron GC content for exons with increasing differential GC content architecture.

(D) Enrichment of FUBP1 binding upstream of the branch point in dependency on intron length and differential GC content. Exons were classified into each intron length groups and then split by GC content architectures (left panel) and vice versa (right panel).

(E) Intron length distribution between kingdoms. Analyses were performed for 174 mammals, 274 non-mammalian vertebrates, 277 invertebrates, 410 fungal species, 94 protozoa, and 145 plants.

(F) Detailed scheme of the mathematical model describing exon definition and splicing for a cassette exon flanked by two constitutive exons. After pre-mRNA synthesis (s), the three exons (indicated by boxes) can be cooperatively and reversibly bound by the pioneering spliceosome subunits U1 and U2 (these are not explicitly displayed in the scheme). Colorless (0) and colored (1) squares represent bound ("defined") and unbound ("undefined") exons, respectively. Red, green, and blue arrows represent binding to and dissociation from exons 1, 2, and 3, respectively, where k_1–k_3 are the corresponding rate constants of binding and k_4–k_6 the rate constants of dissociation. Based on the exon definition patterns (highlighted by red ellipse), splicing decisions towards multiple splice isoforms (inclusion, skipping, first intron retention, second intron retention, full intron retention) are made, and it is assumed that an intron can be excised if the two neighboring exons are defined. For instance, skipping of exon 2 is possible from the state 1_0_1 and occurs with the rate i_{12}. Likewise, splicing of the first intron occurs from the species P1_1_0 and P1_1_1 (rate i_1), and splicing of the second intron from P0_1_1 and P1_1_1 (rate i_2). The inclusion isoform is generated in two steps, i.e., from the subsequent removal of intron 1 and intron 2 in random order. 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, $k_{dr1}+k_{dr2}$: full intron retention).

(G) The intron and exon definition models show similar splicing changes upon *FUBP1* knockout (KO). We simulated the splicing changes upon *FUBP1* KO based on the assumption that FUBP1 affects the rate of spliceosome binding to the 3' splice site of long introns (left panel) or the rate off splicing catalysis across long introns (right panel) as described in detail in the STAR Methods. To account for the heterogeneity of exons in the human genome, we randomly sampled the kinetic parameters of the model 10,000 times to generate an ensemble of 10,000 in silico exons. We then simulated *FUBP1* KO for each in silico exon, assuming that FUBP1 selectively enhances the rate of splicing for long introns, and considered three scenarios reflecting different length configurations of upstream and downstream introns (see STAR Methods for details). The boxplots show the distributions of ΔPSI = PSI(KO) – PSI(control) values for exon (red) and intron definition (blue) across all exons.

(H) Total luminescence and fluorescence measurements were used to estimate the amount of FUBP1 paired with the components of U1 complex (orange), BCL2L1–BAD as a positive control pair (green) and pairs that are not known to interact with each other as negative controls

(gray). Acceptor/donor ratios are similar for all pairs making the cBRET values more comparable to each other.

(I) $^{1}H–^{15}N$ HSQC spectra of the titration of $FUBP1^{A/B}$ with SNRPA up to a molar ratio of 1:1.

(J) $^{1}H–^{15}N$ HSQC spectra of the titration of $PRPF40B^{N\text{-}term\text{-}WW12}$ with $FUBP1^{P\text{-}rich+A/B}$ up to a molar ratio of 1:0.5.

(K) Western blot to verify FUBP1 construct expression after transfection of RPE1 WT and *FUBP1* KO cells.

(L) Capillary electrophoresis of exon inclusion levels of the *MPDZ* minigene after transfection of RPE1 WT and *FUBP1* KO cells with different FUBP1 constructs.

Supplementary Tables

Table S2. Binding affinities and stoichiometries determined by ITC experiments (related to Figures 2D, 2F, S2G–I, and S3H). Experiments were performed for different FUBP1 N-terminal constructs (FUBP1$^{N\text{-}box}$, FUBP1^{N74}) with U2AF2 constructs (U2AF2^{RRM12}, U2AF2$^{linker\text{-}RRM2}$ and U2AF2^{RRM2}) and various FUBP1 KH domain constructs (FUBP1^{KH12}, FUBP1^{KH23}, FUBP1^{KH34} and FUBP1^{KH}) with DNA or RNA.

Analyte	Titrant	N sites	K_D [μM]	Repeats
U2AF2^{RRM12}	FUBP1$^{N\text{-}box}$	0.98 ± 0.17	75.93 ± 2.70	3
U2AF2^{RRM12}	FUBP1^{N74}	0.85 ± 0.07	70.57 ± 2.41	3
U2AF2$^{linker\text{-}RRM2}$	FUBP1$^{N\text{-}box}$	0.91 ± 0.16	78.97 ± 2.16	3
U2AF2^{RRM2}	FUBP1$^{N\text{-}box}$	0.87 ± 0.16	82.03 ± 2.98	3
FUBP1^{KH12}	TTTGTAAAATTTTG	0.78 ± 0.07	4.71 ± 1.45	3
FUBP1^{KH23}	TCTGTAAAATTTGT	0.76 ± 0.09	1.15 ± 0.48	3
FUBP1^{KH34}	TTTTGAAAATCTGT	0.74 ± 0.04	0.87 ± 0.10	3
VPS13D RNA	FUBP1^{KH}	1.38 ± 0.04	0.428 ± 0.062	3

Supplementary References

[S1] Beuth, Barbara, María Flor García-Mayoral, Ian A. Taylor, and Andres Ramos. 2007. "Scaffold-Independent Analysis of RNA-Protein Interactions: The Nova-1 KH3-RNA Complex." *Journal of the American Chemical Society* 129 (33): 10205–10.

[S2] Kang, Hyun-Seo, Carolina Sánchez-Rico, Stefanie Ebersberger, F. X. Reymond Sutandy, Anke Busch, Thomas Welte, Ralf Stehle, et al. 2020. "An Autoinhibitory Intramolecular Interaction Proof-Reads RNA Recognition by the Essential Splicing Factor U2AF2." *Proceedings of the National Academy of Sciences of the United States of America* 117 (13): 7140–49.

[S3] Cukier, Cyprian D., David Hollingworth, Stephen R. Martin, Geoff Kelly, Irene Díaz-Moreno, and Andres Ramos. 2010. "Molecular Basis of FIR-Mediated c-Myc Transcriptional Control." *Nature Structural & Molecular Biology* 17 (9): 1058–64.

[S4] Madeira, Fábio, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, et al. 2019. "The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019." *Nucleic Acids Research* 47 (W1): W636–41.

[S5] Sutandy, F. X. Reymond, Stefanie Ebersberger, Lu Huang, Anke Busch, Maximilian Bach, Hyun-Seo Kang, Jörg Fallmann, et al. 2018. "In Vitro iCLIP-Based Modeling Uncovers How the Splicing Factor U2AF2 Relies on Regulation by Cofactors." *Genome Research* 28 (5): 699–713.

[S6] Seiler, Michael, Shouyong Peng, Anant A. Agrawal, James Palacino, Teng Teng, Ping Zhu, Peter G. Smith, Cancer Genome Atlas Research Network, Silvia Buonamici, and Lihua Yu. 2018. "Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types." *Cell Reports* 23 (1): 282–96.e4.

[S7] Tammer, Luna, Ofir Hameiri, Ifat Keydar, Vanessa Rachel Roy, Asaf Ashkenazy-Titelman, Noélia Custódio, Itay Sason, et al. 2022. "Gene Architecture Directs Splicing Outcome in Separate Nuclear Spatial Regions." *Molecular Cell*. Elsevier.

5. Further investigations into the role of FUBP1 in splicing

5.1. Introduction

In chapter 4, we learned that FUBP1 is a core splicing component regulating the inclusion levels of long introns in a transcriptome-wide manner. We could show that the presence of the FUBP1 splice site is crucial for proper exon inclusion of exons flanked by two long introns. However, we could not determine the mechanisms with which FUBP1 facilitates splicing regulation of long introns. We hypothesize that the presence of FUBP1 alters splicing kinetics and accelerates splicing of long introns. Here, we investigated splicing speed using a nascent RNA-sequencing in both *FUBP1* KD and KO cells. Moreover, we investigated the contribution of the upstream or downstream FUBP1 binding to exon inclusion.

5.2. Materials and Methods

5.2.1. Materials

Table 1: List of Chemicals and materials

Product Name	Company	Product ID
0.2 ml RNase-free PCR tubes	Thermo Fisher	AM12230
1.5 ml reaction tube	Sarstedt	S044598
10 mM dNTPs	New England Biolabs	N0447L
2 ml reaction tube	Eppendorf	30120094
2-Propanol	Carl Roth	9866.1
384-well PCR plate standard	Thermo Fisher	AB-1384
3M sodium acetate pH 5.5	Thermo Fisher	AM9740
4-Thiouridine	Sigma-Aldrich Chemie	T4509-25MG
4-ThioUTP	Jena Bioscience	NU-1156S
Agarose	VWR International	732-2789P
Ampicillin	IMB Media Lab	A-001F20
Biotin-HPDP	Thermo Fisher	15900463
Cell scraper	Sarstedt	833951
Chloroform	Sigma-Aldrich Chemie	25666-100ML
DMEM	Thermo Fisher	11960044
DMEM/F12	Thermo Fisher	11320033
DMSO	Sigma-Aldrich Chemie	472301-500ML
DPBS	Thermo Fisher	14080089
DTT	Sigma-Aldrich Chemie	D9779
EDTA pH 8.0	Thermo Fisher	10135423
ERCC Spike-in Mix	Thermo Fisher	4456740
Ethanol	Sigma-Aldrich Chemie	32205-2.5L-M

Fetal bovine serum	Thermo Fisher	10500056
GeneRuler Ultra Low Range DNA Ladder	Thermo Fisher	10150750
Glutamine	Thermo Fisher	25030081
HS DNA1000 Sample Buffer	Agilent Technologies	5067-5585
HS DNA1000 Screen Tapes	Agilent Technologies	5067-5584
HS RNA Sample Buffer	Agilent Technologies	5067-5580
HS RNA Screen Tapes	Agilent Technologies	5067-5579
Hygromycin B (50 mg/ml)	Thermo Fisher	10453982
LB-Amp agar plates	IMB Media Lab	L-20_P
LB-Luria Medium	IMB Media Lab	L-002RT
Loading Tips	Agilent Technologies	5067-5152
N,N-Dimethylformamide	Sigma-Aldrich Chemie	D4551-250ML
NaCl	Thermo Fisher	AM9760G
Nuclease-free water	Thermo Fisher	10793837
Opti-MEM™ I Reduced Serum Medium	Thermo Fisher	31985054
Penicillin-Streptomycin (10.000 U/ml)	Thermo Fisher	11548876
Phase Lock Heavy Columns	VWR International	733-2478
RNA sample buffer	Agilent Technologies	5067-5577
RNA Screen Tapes	Agilent Technologies	5067-5576
TapeStation tube caps	Agilent Technologies	401425
TapeStation tube strips	Agilent Technologies	401428
TipOne Filter Tip 10/20 µl	Starlab	S1120-3710-C
TipOne Filter Tip 1000 µl	Starlab	S1122-1730-C
TipOne Filter Tip 200 µl	Starlab	S1111-0810
Tissue culture dish optilux 100 mm	Corning	353003
Tissue culture dish optilux 150 mm	Thermo Fisher	353025
Tissue culture plate 6-well	Sarstedt	353046-1
Tris HCl pH 7.0	Thermo Fisher	AM9851
Tris HCl pH 8.0	Thermo Fisher	AM9855G
TRIzol reagent	Thermo Fisher	15596026

Table 2: List of Enzymes

Product Name	Company	Product ID
FuGENE HD Transfection Reagent	Promega	E2312
Gibson Assembly Master Mix	IMB Protein Production	n.a.
Luminaris Color HiGreen Low ROX qPCR Master Mix	Thermo Fisher	13515260
OneTaq DNA Polymerase	New England Biolabs	M0480L
Phusion HighFidelity Polymerase	New England Biolabs	M0530S
Q5 DNA Polymerase	New England Biolabs	M0491L
SuperScript III Reverse Transcriptase	Thermo Fisher	18080044
Trypsin	Thermo Fisher	15090-046
TURBO DNase (2U/l)	Thermo Fisher	10722687

Table 3: List of Devices

Product Name	Company	Product ID
Applied Biosystems® ViiA™ 7 Real-Time PCR System	Thermo Fisher	4453536
ChemiDoc XRS+ Imaging System	Bio-Rad	1708265
CO2 Incubator Heracell	Thermo Fisher	41787784
Electrophoresis chamber	Thermo Fisher	EI0002
Heraeus Multifuge X3R	Thermo Fisher	75004515
Heraeus Pico 21 Microcentrifuge	Thermo Fisher	75002553
Multitron 2 Shaking Incubator	Infors HT	S-000118319-003
NGS Quest Plus chromatography system	Bio-Rad	7880003
PCR Cycler Biometra T Advanced	Analytic Jena	3905270
Qubit 2.0 Fluorometer	Thermo Fisher	75004515
Safe 2020 Biological Safety Cabinet	Thermo Fisher	10462614
Spectophotometer DS-11	DeNovix	S-06028
TapeStation 4200	Agilent	G2964-90003
ThermoMixer F1.5	Eppendorf	5384000012

Table 4: List of Plasmids

Internal No	Plasmid name	Creator
pAB010	pcDNA5-mpdz-Minigene	[redacted]
pMMM039	pcDNA5_mpdz_Δintron1	Miriam Mulorz
pMMM040	pcDNA5_mpdz_Δintron1+ΔBS1	
pMMM041	pcDNA5_mpdz_Δintron2	
pMMM042	pcDNA5_mpdz_Δintron2+ΔBS2	
pMMM043	pcDNA5_mpdz_Δintron	
pMMM044	pcDNA5_mpdz_Δintron+ΔBS	
pMMM049	pcDNA5_mpdz_ΔBS1	
pMMM050	pcDNA5_mpdz_ΔBS2	
pMMM051	pcDNA5_mpdz_ΔBS	

Table 5: List of Oligos

Purpose	Number	Name	Sequence
Spike-in generation primers	oMMM1	Spike-in1 fw	TAATACGACTCACTATAGGGTGCTTTAACAAGAGGAAATTGTGT
	oMMM2	Spike-in1 rv	CCATCTTGTTTATAAAATCCTAATTACTC
	oMMM3	Spike-in2 fw	TAATACGACTCACTATAGGGGGCACAAGTTGCTGAAGTTGC
	oMMM4	Spike-in2 rv	TCTGCTGTAATCTCAGCTCC
	oMMM5	Spike-in3 fw	TAATACGACTCACTATAGGGTTTCGACGTTTTGAAGGAGGG
	oMMM6	Spike-in3 rv	GTACCCGGGAAAATCCTAGTTC

	oMMM7	Spike-in4 fw	TAATACGACTCACTATAGGGACTGTCCTTTCATCCATAAGCGG
	oMMM8	Spike-in4 rv	CGCACGCCGAATGATGAAACG
	oMMM9	Spike-in5 fw	TAATACGACTCACTATAGGGGATGTCCTTGGACGGGGT
	oMMM10	Spike-in5 rv	GCTTTCGGAGCAAATCGCG
	oMMM11	Spike-in6 fw	TAATACGACTCACTATAGGGCCAGATTACTTCCATTTCCGCC
	oMMM12	Spike-in6 rv	GGGTAAAACGCAAGCACCG
Spike-in qPCR primers	oMMM13	qSpike-in1 fw	ACAATTCCAAATAGCGACCACATCA
	oMMM14	qSpike-in1 rv	TACCTCAACCCTTCCAGTGTCTAAG
	oMMM15	qSpike-in2 fw	AGACTGGCATTCCCGTGATA
	oMMM16	qSpike-in2 rv	GCTAAAACCCCTGCCTGCAA
	oMMM17	qSpike-in3 fw	CCGAGTTCGCCTTACTGCTC
	oMMM18	qSpike-in3 rv	AATCGATCGGAATCACGCCG
	oMMM19	qSpike-in4 fw	CATAAGCGGAGAAAGAGGGAATGAC
	oMMM20	qSpike-in4 rv	GCTAAATAGAGAGCATCCACACCTC
	oMMM21	qSpike-in5 fw	CGTTAATGCAGAGGCTAAGGACAAT
	oMMM22	qSpike-in5 rv	GATCGTTACAAACCCACTACGTGTC
	oMMM23	qSpike-in6 fw	GTCCTGATTTACTGGACTCGCAAC
	oMMM24	qSpike-in6 rv	TCTGTATAAGGTGATCGCAGGTTGT
Cloning primers for *MPDZ* minigenes	oMMM96	mpdz_short_in1_F	ATTCTATGTTGAAGTCATATCG
	oMMM97	mpdz_short_in1_R	TAAATCTTTATACGAATGGAG
	oMMM98	mpdz_short_in2_F	TTAGCAGTGTCTGATGAG
	oMMM99	mpdz_short_in2_R	TTGTGAGAACTAAGTTGG
	oMMM100	mpdz_lin_F	TAATCTCGAGTGAAAATGGGAAGAAAAACC
	oMMM101	mpdz_lin_R	TTTAACTACAATGCAGCAGAACAAGGATAAATAG
	oMMM104	mpdz_short_in1_with_FUBP1_F	AGGGATTCTCATCCAACAG
	oMMM105	mpdz_short_in1_with_FUBP1_R	TAAATCTTTATACGAATGGAGC
	oMMM106	mpdz_short_in1_no_FUBP1_F	CTTTCTCTTCTCGCCCCC

	oMMM107	mpdz_short_in1_no_FUBP1_R	TAAATCTTTATACGAATGGAGCTAG
	oMMM108	mpdz_short_in2_no_FUBP1_F	TATGTAATAGAACAGAATTCCC
	oMMM125	G08_MM_q5_mpdz_no_BS_in1_R	CTGTTGGATGAGAATCCCT
	oMMM127	G08_MM_q5_mpdz_no_BS_in2_R2	CTCATCAGACACTGCTAA
	oMMM128	G08_MM_q5_mpdz_no_BS_in2_R2	CTCATCAGACACTGCTAA
RT-PCR for *MPDZ* minigenes	oABu061	f_MPDZ	GGGACTGTGAGAATAGGAGTTGC
	oLa039	RON_RTvalidation_minigene_os67_r	GCAACTAGAAGGCACAGTCG
RT-PCR for *FAM126A*	oABu071	f_FAM126A	ACCAGCATGTCAATAAGGGGTCA
	oABu072	r_FAM126A	TGGTACTAGACGCAGCCCTGGA
siRNA for Ctrl KD	sR39	non-target siRNA 1	UGGUUUACAUGUCGACUAA
	sR40	non-target siRNA 2	UGGUUUACAUGUUGUGUGA
	sR41	non-target siRNA 3	UGGUUUACAUGUUUUCUGA
	sR42	non-target siRNA 4	UGGUUUACAUGUUUUCCUA
siRNA for *FUBP1* KD	sR36	FUBP1 siRNA sense	CAGCAAAGCAGAUCUGUAA[DT][DT]
		FUBP1 siRNA antisense	UUACAGAUCUGCUUUGCUG[DT][DT]
	sR43	FUBP1 siRNA sense	CUGGAACACCUGAAUCUGU[DT][DT]
		FUBP1 siRNA antisense	ACAGAUUCAGGUGUUCCAG[DT][DT]
	sR75	FUBP1 siRNA sense	GGCAGGAACGGAUCCAAAU[DT][DT]
		FUBP1 siRNA antisense	AUUUGGAUCCGUUCCUGCC[DT][DT]

Table 6: List of Buffers

Purpose	Buffer	Composition	Details
nascent RNA-seq	4sU stock solution	100 mM 4-Thiouridine	Dissolve in DMSO, aliquot in light-protected tubes and store at -20 °C.
	Biotinylation Buffer	100 mM Tris HCl pH7.4 10 mM EDTA pH 8.0	Store at 4 °C.
	HPDP stock solution	1 mg/ml HPDP	Dissolve in 1 ml DMF, aliquot and store at -20 °C.
	Wash buffer	100 mM Tris HCl pH7.4 10 mM EDTA pH 8.0 1 NaCl 0.1% (v/v) Tween20	Store at 4 °C.

5.2.2. Methods

5.2.2.1. RPE1 Cell Culture

Three hTERT-RPE1 cell lines "*FUBP1* WT", "*FUBP1 N-box*^{mut}" and "*FUBP1* KO" were derived from the hTERT-RPE1 NatNeo Cas9 Mono Puro sens cell line. This cell line was a generous gift from the [redacted] lab at the [redacted], and was modified from original hTERT RPE1 cells (ATCC, CRL-4000). *FUBP1 N-box*^{mut} and *FUBP1* KO were gained by CRIPSR/Cas9 editing as described above in chapter 4. After gene editing, all cell lines were kept in neomycin-free conditions. The cell lines were grown and maintained in Dulbecco's modified Eagle's medium:Nutrient Mixture F-12 (DMEM/F-12), supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, 1% penicillin–streptomycin, and 20 µg/ml hygromycin B. Cultivation took place by seeding 1×10^6 cells in a 10-cm cell culture dish and incubating at 37 °C with 5% CO_2. For subcultivation, cells were detached from the dish with 3 ml of 0.1% trypsin every 2–3 days for 20 passages. After that, new cells were thawed from aliquots containing 1×10^6 cells in 1 ml of growth medium, supplemented with 10% DMSO and 50% FBS.

5.2.2.2. HeLa cell culture

HeLa cells were kept in DMEM, supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, 1% penicillin–streptomycin. Cultivation, subcultivation and thawing of aliquoted cells was performed as described above for 5.2.2.1, RPE1 Cell Culture.

5.2.2.3. siRNA-mediated Knockdowns

For the 4sU-sequencing, siRNA-mediated *FUBP1* and Ctrl Knockdowns (KD) were performed. One day prior to siRNA treatment, 2.0×10^5 HeLa cells were seeded in 1 well of a 6-well plate, or 1.0×10^6 cells in a 10 cm dish. 24h later, 30 nM total siRNA was transfected. We used the siRNA in Table 5. For 1 well of a 6-well plate, 3 µl of a 10 µM siRNA Mix was diluted in 100 µl OptiMEM. Likewise, 3 µl RNAimax were diluted in 100 µl OptiMEM. The two solutions were mixed and incubated for 20 min before carefully pipetted onto the cells. For a 10 cm dish, 15 µl of a 10 µM siRNA mix and 15 µl RNAimax were used in 500 µl OptiMEM. The transfection was incubated for 48h at 37 °C, then the success of the KD was measured by RT-PCR.

5.2.2.4. Bacterial culture and vector cloning

All cloning expression and vector expressions took place in DH5α bacterial cells. Bacteria was grown in LB medium supplemented with ampicillin (LB-Amp) and grown overnight at 37 °C, 300 rpm. For cloning and starter cultures of re-transformation, 2 ml LB-Amp medium were used, while for vector amplification, 1 ml of starter culture was diluted in 25 ml LB-Amp. Plasmids were isolated using the ZymoPURE Plasmid Miniprep Kit or the QIAGEN Plasmid Plus Midi Kit. Elution was achieved with 20-200 µl nuclease-free H_2O.

We mutated the *MPDZ* minigene and the FUBP1 expression vector in various ways. The *MPDZ* minigene was obtained from [redacted], who isolated the *MPDZ* minigene from HeLa gDNA extracts by amplification of chr9:13,183,353-13,189,041 with Phusion HighFidelity Polymerase before the PCR fragment was cloned into a pcDNA5 vector by Gibson assembly. *MPDZ* introns were shortened by site-directed mutagenesis using a Q5 Site-Directed Mutagenesis Kit, 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, as described in chapter 4. In addition, we created four more clones that have not been discussed in chapter 4. $MPDZ^{\Delta intron1}$ only lacks chr9:13,183,736-13,186,120 while $MPDZ^{\Delta intron2}$ has a deletion in chr9:13,186,637-13,188,633. Likewise, $MPDZ^{\Delta BS1}$ is missing chr9:13,186,637-13,188,633 and $MPDZ^{\Delta BS2}$ lacks chr9:13,183,632-13,186,120.

5.2.2.5. *Transfection of minigenes*

To perform the minigene splicing assays, *MPDZ* minigenes were transfected 1×10^5 RPE1 cells. The cells were seeded 24h prior to transfection. Then, 2 µg of DNA were mixed with 100 µl OptiMEM. Then, 6.4 µl FuGENE were carefully added to the diluted DNA to create a 1:3.2 DNA:FuGENE mix. After 10 min of inclubation, the transfection mix was added to the cells and inclubated for 24h. The splicing of the transfected *MPDZ* minigene was evaluated by RT-PCR.

5.2.2.6. *RT-PCR for* FUBP1 *KD, minigene splicing assay and complementation*

The semi-quantitative reverse transcription followed by polymerase chain reaction (RT-PCR) was performed to evaluate the siRNA-mediated *FUBP1* KD, but also to analyze the *MPDZ* minigene splicing assay. After successful transfection or KD, cells were washed twice with ice-cold PBS and harvested in 500 µl PBS using a cell scraper and cells were pelleted by centrifugation at 1 000 x g, 1 min, 4 °C. RNA isolation was performed using the RNeasy PLUS Mini Kit according to manufacturer's instructions. 300 ng of total RNA was used for subsequent RT with the RevertAid First Strand cDNA Synthesis Kit. All PCRs were performed using 0.5 µl template and the One-Taq Polymerase according to the manufacturers protocol. The conditions used for the *FAM126A* PCR (KD validation) and the *MPDZ* PCR (minigene splicing assay) are indicated in Table 7 and Table 8, respectively. The splicing isoforms were analyzed by capillary electrophoresis using the Agilent Tape Station and the HS DNA1000 Screen Tapes according to manufacturer's protocol. The significance was determined by a Student's t-test followed by a Benjamini-Hochberg correction for multiple testing.

Table 7: Conditions for *FAM126A* PCR

Cycles	Step	Temperature	Time
	Initiation	94 °C	30 sec
28 cycles	Denaturation	94 °C	30 sec
	Annealing	62 °C	30 sec
	Elongation	68 °C	1:20 min
	Final Extension	68 °C	5 min
	Pause	16 °C	hold

Table 8: Conditions for *MPDZ* PCR

Cycles	Step	Temperature	Time
	Initiation	94 °C	30 sec
21 cycles	Denaturation	94 °C	30 sec
	Annealing	52 °C	30 sec
	Elongation	68 °C	40 sec
	Final Extension	68 °C	5 min
	Pause	16 °C	hold

5.2.2.7. Spike-in generation

To normalize nascent RNA-sequencings, we add both 4sU-labeled and non-labeled spike-ins to the total RNA. The spike-ins were generated from six different ERCC Spike-in Mix, which was obtained in a 1:1000 dilution from the [redacted]. The whole mix was reverse transcribed using the SuperScript III Kit according to manufacturer's protocol. Here, 1 µl of RNA dilution was used. The resulting cDNA was stored at -20 °C. Then, six spike-ins were chosen to be used in subsequent spike-in generation and nascent RNA-seq according to Schwalb et al. Protocol for the spike-in characteristics and generation has been described in the TT-seq bench protocol from Gressel and colleagues (Gressel, Lidschreiber, and Cramer 2019). Each spike-in was PCR amplified using the Q5 DNA Polymerase and sequence-specific primer, harboring a T7 overhang for later *in vitro* transcription. The utilized primers can also be found in Table 5. For a 50 µl reaction, 5 µl of cDNA was used. The success of the amplification was verified by gel electrophoresis using an agarose gel. The PCR reaction was purified using the QIAquick PCR Purification Kit. The resulting RNA spike-ins were *in vitro* transcribed with the HiScribe T7 High Yield RNA Synthesis Kit according to the manufacturer's protocol. For three spike-ins (#1, #3, and #5), the *in vitro* transcription took place in presence of 10% 4sUTP. For the remaining three spike-ins (#2, #4, and #6), the unlabeled UTP provided with the HiScribe T7 High Yield RNA Synthesis Kit was used. After *in vitro* transcription, the input DNA was digested with TURBO DNase and purified with the RNeasy MinElute Cleanup Kit and eluted in 15 µl of water. Concentration was measured with Qubit and RNA spikes were pooled to 10 ng/µl per spike-in RNA. Spike-in pool was aliquoted and stored at -80 °C.

5.2.2.8. Nascent RNA-sequencing

This protocol is based on both Rädle and colleagues and Schwalb and colleagues (Rädle et al. 2013; Schwalb et al. 2016). Nascent RNA-sequencing was performed in HeLa and RPE1 cells. In HeLa cells, a *FUBP1* or control (Ctrl) Knockdown (KD) carried out prior to the experiment. $2.4x10^6$ cells were seeded in a 10 cm culture dish and treated with siRNA as described in 5.2.2.3. Incubation took place for 48h before nascent RNA-seq. To achieve 4sU labeling, growth medium was aspired from the cells and 5 ml of the medium was added to with 200 µM 4sU in DMSO (10 µl of the 100 mM 4sU stock solution). After brief and thorough mixing, the 4sU-containing medium was added back to the cells and incubation took place for 60 min at 37 °C. The reaction was quenched by again aspiring the medium and quickly adding 2 ml of TRIzol. After 3 min of incubation, the lysed cells were collected from the culture dish and distributed in 1 ml aliquots in 1.5 ml reaction tubes. RPE1 were processed similarly with some alterations. $3.5x10^6$ cells were seeded in a 15 cm culture dish, 20 µl 100 mM 4sU stock solution was used in 10 ml of growth medium, and the cells were lysed with 3 ml of TRIzol. The samples can be frozen at this time point and thawed by heating the samples to 65 °C for 5 min.

Next, RNA extraction was performed. Here, 200 µl of Chloroform was added to 1 ml TRIzol lysis reagent and mixed vigorously. After shaking vigorously, samples were incubated on ice for 3 min until the phases separate. Now, samples were centrifuged at 4 °C, 13 000 x g for 15 min. The upper aqueous phase containing the RNA was transferred into a new 1.5 ml reaction tube and 1/10 volume NaOAc and 1 volume isopropanol was added. RNA precipitation was achieved by incubating the samples at -20 °C for 1h. Subsequently, cells were centrifuged for 10 min at 4 °C at 20 000 x g, and the resulting pellet was washed with 75% Ethanol and centrifuged again for 10 min at 4 °C at 20 000 x g. After the Ethanol was aspired, RNA was dissolved in 200 µl per sample. The integrity of the RNA was examined by capillary electrophoresis using the Agilent Tape Station and the RNA Screen Tapes.

For biotinylation of the 4sU, RNA was measured with Qubit according to manufacturer's protocol and diluted to 30 ng in 210 nuclease-free H_2O and 30 µl 10x biotinylation buffer, 30 µl DMF, 30 µl Biotin-HPDP and 0.3 pg of spike-in mix was added. Biotinylation took place for 1.5 h at room temperature with 300 rpm. Subsequently, RNA was isolated by adding an equal volume, here 300 µl, of chloroform and mixing vigorously, followed by 3 min of incubation until phases start to separate. The sample was centrifuged at 20 000 x g for 5 min at 4 °C and the upper aqueous phase was transferred into a new tube. Here, again an equal volume of chloroform was added, mixed and then transferred to a Phase Lock Gel Heavy tube before centrifuging and transferring again. RNA was precipitated by adding 1/10 volume of 5 M NaCl and an equal volume of isopropanol. To obtain an RNA pellet, RNA was centrifuged for 20 min at 20 000 x g at 4 °C. Supernatant was discarded

and the pellet was washed with 75% ethanol and centrifuged again for 10 min at 20 000 x g at 4 °C. Supernatant was removed and RNA was eluted in 30 µl nuclease-free H_2O. The integrity of the biotinylated RNA was examined by capillary electrophoresis using the Agilent Tape Station and the RNA Screen Tapes.

Isolation of nascent RNA was achieved with the µMACS Streptavidin Kit. The biotinylated samples were denatured at 65 °C for 10 minutes and then immediately placed on ice. Then, the 100 µl of streptavidin beads were added to the samples and incubated for 15 min at room temperature with 300 rpm. Meanwhile the Miltenyi columns were pre-equilibrated with 1 ml room temperature washing buffer. After incubation, the streptavidin-RNA mix was added to the column. The sample was then washed three times with 65 °C washing buffer, then three times with room temperature washing buffer. After washing, the RNA was released from the streptavidin beads by eluting with 2 x 100 µl 0.1 M DTT. The now obtained nascent RNA was cleaned with the RNeasy MinElute Cleanup Kit. The integrity of the nascent RNA was examined by capillary electrophoresis using the Agilent Tape Station and the HS RNA Screen Tapes. Concentration was measured with the DeNovix spectrophotometer. Library preparation was performed using the NuGEN Ovation RNA solo Kit. Next Generation Sequencing was performed as a 150 nt single end high output read on 1 flow cell.

The 4sU-seq libraries for both experiments were sequenced on an Illumina NextSeq 500 sequencing machine as 145 nt single-end reads. The index read included 8 nt UMIs (unique molecular identifiers) in addition to the sample index. UMIs were extracted as a second FASTQ file for each sample and later added to the read headers of the 145 nt reads using UMI-tools (v1.0.0) (Smith, Heger, and Sudbery 2017). Basic quality controls were performed for all 4sU-seq samples with FastQC (v0.11.8) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Reads were mapped using STAR (v2.7.3a) (Dobin et al. 2013) and allowing up to 4% of the mapped bases to be mismatched (--outFilterMismatchNoverLmax 0.04 --outFilterMismatchNmax 999) and a splice junction overhang (--sjdbOverhang) of 144 nt. Genome assembly and annotation of all reference chromosomes in GENCODE (Frankish et al. 2019) release 31 were used during mapping in combination with the sequences of all ERCC spike-in genes. Subsequently, multi-mapping reads were removed using Samtools (v1.9) (Danecek et al. 2021). Making use of the UMIs, duplicates of the remaining uniquely mapped reads were removed using UMI-tools (v1.0.0) (Smith, Heger, and Sudbery 2017).

The SJ file from STAR serves as basis for the evaluation of used introns by means of junction reads. The SJ file contains high confidence splice junctions. Introns were analyzed when being present in the SJ file in 3 out of 4 replicates resulting in about 110 000 – 125 000 analyzed introns per condition (control, KD, WT, MUT, KO). Introns were further filtered for overlap with Gencode release 31

annotation which in turn was filtered for protein coding genes and transcripts, TSL < 4 and level < 3, resulting in about 100,000 introns (control, *FUBP1* KD) and 90,000 introns (WT, *FUBP1 N-box*^{mut}, *FUBP1* KO). Non-split reads are reads that either completely overlap the last 5 nt of the intron and the first 5 nt of the exon or, alternatively, the last 5 nt of the exon and the first 5 nt of the intron. Split reads are reads that have a junction at the respective intron. Only introns with read counts in the top 20% quantile range in either average ctrl non-split/split reads or average MUT/KO non-split/split reads are taken into account. The number of introns in each intron length group is about 5,000. Intron half-life analysis is done as follows: intron counts are compared between conditions. Normalization for expression strength is done by dividing intron counts by the number of reads mapping to all exons in that gene. Only introns that have counts in the top 20% quantile range in either average ctrl/WT intron/exon reads or average *FUBP1* KD/*FUBP1 N-box*^{mut}/*FUBP1* KO intron/exon reads.

5.2.2.9. qPCR for spike-in validation

To observe whether the 4sU-labeled RNA spike-ins were pulled down more efficiently compared to the non-labeled RNA spike-ins, a RT-qPCR was performed. Therefore, the 1 µl of nascent RNA product was reverse transcribed using the SuperScript III Reverse Transcriptase according to manufacturer's protocol. The cDNA Mix was then diluted 1:10 prior to using it in the qPCR. 2 µl of diluted cDNA was added to a well of a 384-well plate and mixed with 10 µl Luminaris Color HiGreen Low ROX qPCR Master Mix, 0.6 µl of each reverse and forward primer (Table 5) and filled with 6.8 µl of H_2O, resulting in a 20 µl mix. The qPCR program is stated in Table 9. Ct values were normalized to spike-in #1 and fold enrichment was calculated. Error bars indicate standard deviation from 3 technical replicates.

Table 9: qPCR cycle conditions

Cycles	Step	Temperature	Time
	UDG pre-treatment	50 °C	2 min
	Initiation	95 °C	10 min
40 cycles	Denaturation	95 °C	15 sec
	Annealing/Extension	60 °C	60 sec

5.3. Results

5.3.1. Nascent RNA-sequencing was performed to investigate splicing speed

5.3.1.1. Establishing and optimizing the nascent RNA sequencing in HeLa cells

To examine splicing speed in the presence and absence of FUBP1, we performed a nascent RNA sequencing. In brief, cells were labeled for 1h with 4-thiouridine (4sU), which incorporates into the RNA instead of uridine. Subsequently, the reaction was quenched by TRIzol and total RNA was isolated. The RNA is biotinylated with biotin-HPDP, a molecule that specifically biotinylates 4sU. With a streptavidin pulldown, the 4sU-labeled, biotinylated RNA was isolated, resulting in nascent RNA that has been transcribed within the 60 min of labeling (**Figure 5**A). Here, we use the protocol of Rädle and colleagues (Rädle et al. 2013). We also used both 4sU-labeled and unlabeled RNA spike-ins derived from Schwalb and colleagues, to investigate the success of the pulldown and to normalize the nascent RNA-seq (see 5.2.2, Methods) (Schwalb et al. 2016).

To investigate whether the protocol leads to an enrichment of nascent RNA, we controlled our experiment for several factors. First, we determined the amount of 4sU necessary for the experiment. Because 4sU has reported to be cytotoxic, we tested the effect of high (800 µM) and low (200 µM) concentration of 4sU on WT HeLa cells. The cells labeled with high 4sU concentration showed significant cell death in 1h (not shown). Therefore, we decided to perform the following experiment with low concentration of 4sU (200 µM).

Next, we estimated the amount of spike-ins needed for the experiment. The protocol of Schwalb and colleagues (Schwalb et al. 2016) suggest adding spike-ins before RNA isolation when performing time course experiments. However, to normalize pre-mRNA amounts between *FUBP1* KO and WT conditions, we decided to add spike-ins after RNA isolation and before biotinylation. Therefore, we decreased the amount first to 1 pg of spike-in. We performed a trial run and though we were able to isolate nascent RNA, we could not amplify our 4sU-labeled spike-ins by RT-qPCR, indicating that the input amount was too low. Therefore, we increased the RNA spike-in input to 10 pg. Thereafter, we could indeed observe an enrichment of 4sU-labeled spike-ins compared to the unlabeled spike-ins, validating that we indeed enrich nascent RNA (**Figure 5**B).

Finally, to investigate whether the pulldown of nascent RNA was successful, we examined the presence and integrity of the RNA at three different stages. We observed total RNA, biotinylated RNA and the final product, the nascent RNA (**Figure 5**C). We expected high quality RNA for total and biotinylated RNA, which is characterized by a high percentage of ribosomal subunits in the RNA pool. The ratio of 28S to 18S is measured by capillary electrophoresis and displayed in the RNA integrity value (RIN). As expected, we obtain high RIN values between 8.6 and 10.0. For the nascent RNA, the RIN values do not apply, as the ribosomal RNA (rRNA) amount is drastically decreased. Additionally, the nascent RNA samples contain RNA species with high molecular weight, further

distorting the measurement of RIN values. Indeed, we could determine an increase for unprocessed RNA (> 6 000 nt). Taking together, by RT-qPCR and capillary electrophoresis, we could show that the nascent RNA pulldown is indeed able to enrich newly transcribed RNA.

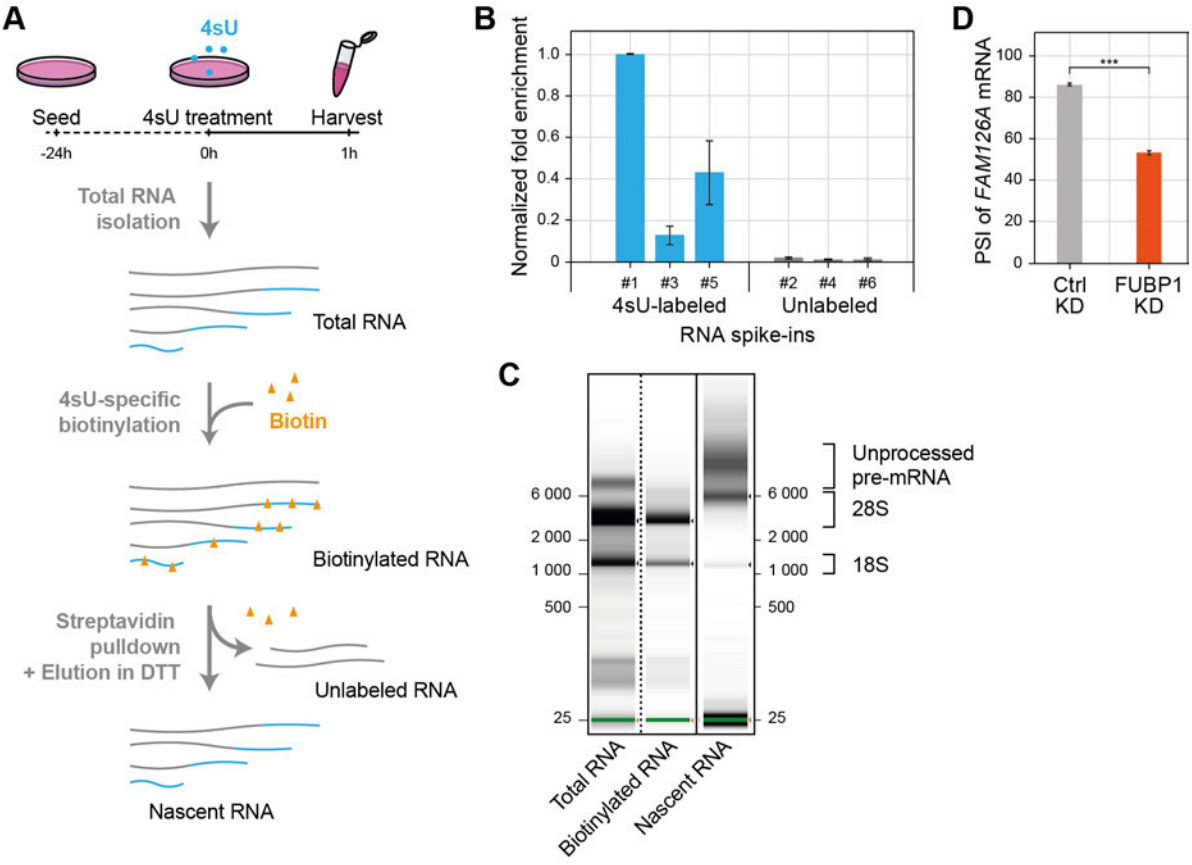


Figure 5: Nascent RNA-seq establishing and quality control. (A) Workflow of nascent RNA-seq. 1h post-treatment, cells are harvested and after total RNA isolation and biotinylation, the labeled nascent RNA is separated from unlabeled RNA using streptavidin beads. (B) RT-qPCR of RNA spike-ins after nascent RNA-seq protocol using 1 pg of RNA spike-in mix per 1 µg of total RNA. 4sU-labeled spike-ins (blue) are enriched compared to unlabeled spike-ins (gray). (C) Example of capillary electrophoresis following nascent RNA-seq protocol. Compared to total and biotinylated RNA samples, there is an enrichment of long, unprocessed RNA species in the nascent RNA sample. (D) Validation of *FUBP1* KD in HeLa cells. RT-PCR of the FUBP1 target *FAM126A* shows FUBP1-specific skipping of alternative *FAM126A* exon 11.

5.3.1.2. Impairment of *FUBP1* does not alter splicing pattern 1h post-labelling

To investigate splicing changes upon *FUBP1* depletion, we performed two nascent RNA-sequencing experiments. First, we performed a control and *FUBP1* Knockdown (KD) in HeLa cells prior to the experiment. The knockdown efficiency was confirmed by RT-PCR of a known FUBP1 target, *FAM126A* mRNA. In HeLa RNA-seq, *FUBP1* KD has shown to reduce inclusion levels of *FAM126A* exon 11. Indeed, we could see a reduction of *FAM126A* exon 11 inclusion levels from 86% to 53%. (**Figure 5**D). The RPE1 cell line did not need any further validation, as we already explored the cell line in chapter 4.

After Next-Generation-Sequencing, we divided the obtained splicing events according to their intron length, as has been done in chapter 4. Now, we compared splicing between WT or Ctrl KD and their corresponding *FUBP1*-depletion for all intron length bins. To investigate splicing velocity, we compared split-reads to non-split reads (**Figure 6**A). In HeLa Ctrl vs *FUBP1* KD, we saw that generally, in *FUBP1* KD condition, the introns were spliced slower than in the Ctrl KD (**Figure 6**B, left). The same held true for RPE1 WT compared to *FUBP1 N-box*^{mut} (**Figure 6**B, middle). Interestingly, for RPE1 WT vs *FUBP1* KO, we observed the opposite effect (**Figure 6**B, right). Here, splicing in *FUBP1* KO was faster than in *FUBP1* WT.

Further, we examined the longevity of introns with different lengths. Therefore, we compared the relative amount of an intron to its neighboring exons. Again, *FUBP1* KD and *FUBP1 N-box*^{mut} exhibit longer half-life of introns compared to Ctrl KD and RPE1 WT, respectively. In addition, very short and very long introns seemed to persist the longest, when FUBP1 is reduced. However, we again found the opposite effect within the *FUBP1* KO cells (**Figure 7**).

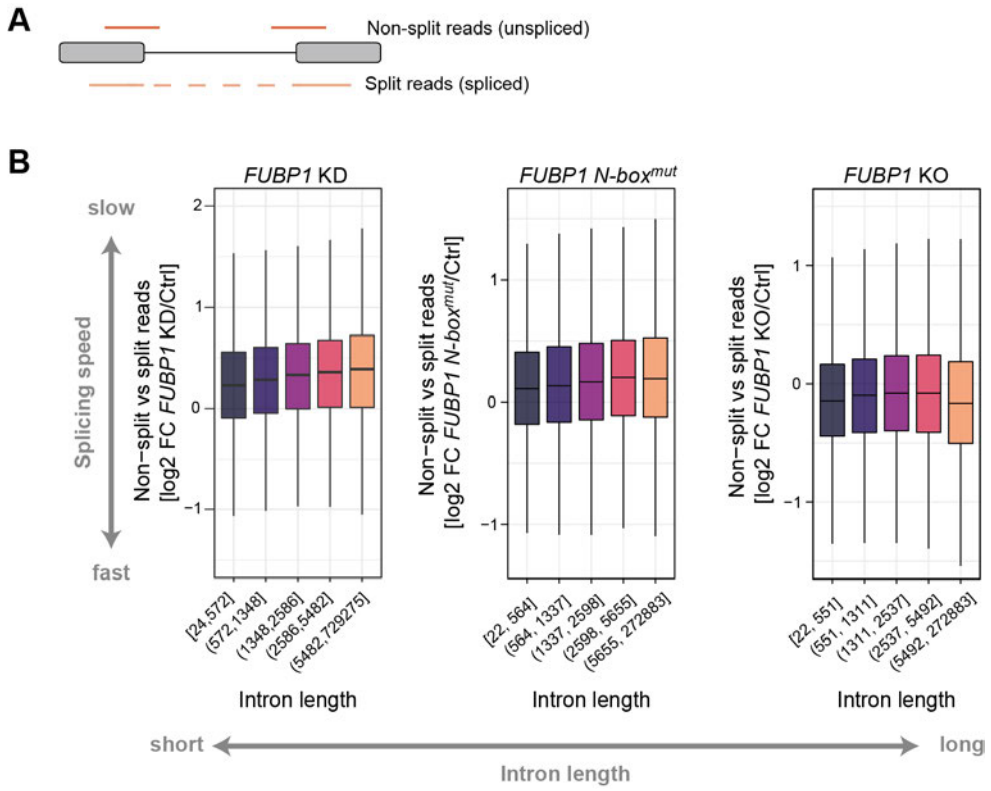


Figure 6: Splicing speed analyzed by nascent RNA-seq. (A) To analyze splicing speed, non-split reads were compared to split reads. This represents non-spliced versus spliced mRNA. (B) Splicing speed analysis for *FUBP1* KD, *FUBP1 N-box*^{mut} and *FUBP1* KO compared to the respective controls. Introns were divided into five bins according to intron length. Though in *FUBP1* KD and *FUBP1 N-box*^{mut}, splicing speed decreases for long introns, this cannot be reproduced in *FUBP1* KO.

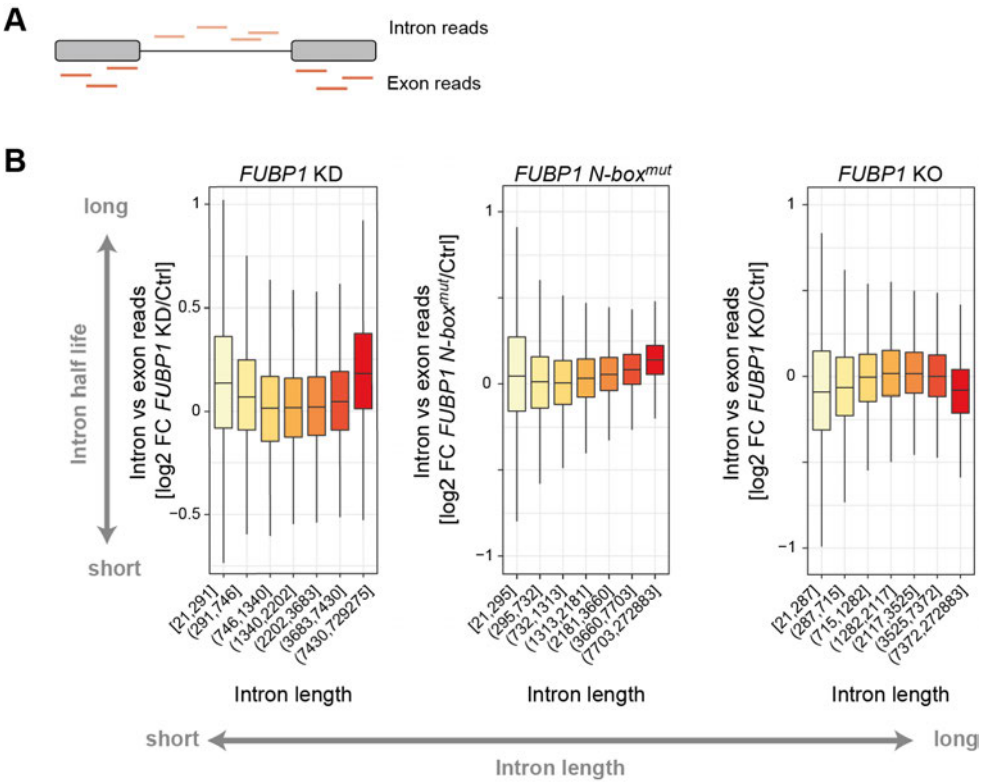


Figure 7: Relative intron half-life analyzed by nascent RNA-seq. (A) To analyze relative intron half-life, intron reads were compared to exon reads. This represents relative amount of introns versus exons. (B) Relative intron half-life analysis for *FUBP1* KD, *FUBP1 N-box*^{mut} and *FUBP1* KO compared to the respective controls. Introns were divided into seven bins according to intron length. Though in *FUBP1* KD and *FUBP1 N-box*^{mut}, long introns live longer in *FUBP1*-depleted condition compared to the respective control, again, this cannot be reproduced in *FUBP1* KO.

As we could clearly observe that FUBP1 is crucial for the correct inclusion of exons flanked by long introns, we expected to see slower splicing of long introns when FUBP1 is missing. However, our data is not conclusive on this matter. Therefore, we cannot make a statement about how FUBP1 influences splicing. A number of explanations for these results have been described in the Discussion.

5.3.2.Mutation of single FUBP1 binding sites in MPDZ minigene leads to interesting results

To examine how FUBP1 influences splicing of exons flanked by long introns, we generated an *MPDZ* minigene, comprising of three exons, with exon 2 being alternatively spliced. The alternative exon is separated from the other exons by two long introns (**Figure 8**A, Deletions). As described in chapter 4, we investigated how the inclusion levels of exon 2 change when we either deplete the FUBP1 binding sites, shorten the introns, or both. We found that the inclusion of exon 2 depends on the presence of FUBP1, and the mere deletion of the FUBP1 binding sites lead to a drastic decrease in inclusion levels. These results are reiterated in **Figure 8**B, lane 1 and lane 8-10.

However, we only examined minigenes in which both introns or both binding sites were altered simultaneously. Now, we aimed to understand whether FUBP1 acts on either the up- or the

downstream intron to regulate inclusion levels. In addition, we compared whether the binding site or the intron length affect the splicing decisions. Therefore, we altered the intron features in only the first or second intron (**Figure 8**A, Further deletions). Subsequently, semi-quantitative RT-PCR was performed to display relative inclusion rates.

First, we determined the impact of the upstream FUBP1 binding site on exon 2 inclusion (*MPDZ*$^{\Delta BS1}$, *MPDZ* $^{\Delta intron1+\Delta BS1}$ and *MPDZ*$^{\Delta BS}$). Notably, the mutants containing the deletion of the upstream binding site (*MPDZ*$^{\Delta BS1}$, *MPDZ* $^{\Delta intron1+\Delta BS1}$ and *MPDZ*$^{\Delta BS}$) had the strongest negative impact on inclusion levels. They showed inclusion levels of approx. 8% (**Figure 8**B, lanes 2, 4 and 9, respectively). Interestingly, this is also true for *MPDZ* $^{\Delta intron1+\Delta BS1}$, in which the intron is shortened as well. This indicates that the intron length itself does not regulate exon inclusion, but the relatively strong FUBP1 binding site in long introns does. In addition, all cell lines displayed comparable inclusion levels, indicating that the dependency on FUBP1 was completely lost. This shows that FUBP1 binding to the FUBP1 binding site within the upstream intron is crucial for downstream intron inclusion.

Next, we examined the splicing pattern changes upon deletions in the downstream intron. Interestingly, they showed a contrasting effect to alterations in the upstream intron. While the deletion of the upstream FUBP1 binding site in *MPDZ*$^{\Delta BS1}$ and *MPDZ* $^{\Delta intron1+\Delta BS1}$ clearly decreased exon 2 inclusion (**Figure 8**B, lanes 2 and 4), deletion of the downstream FUBP1 binding site in *MPDZ*$^{\Delta BS2}$ and *MPDZ* $^{\Delta intron2+\Delta BS2}$ resulted in increased exon 2 inclusion (**Figure 8**B, lanes 5 and 7). The strongest elevation of inclusion could be seen in *MPDZ*$^{\Delta BS2}$, in which only the downstream FUBP1 binding site is deleted, but the length of the intron is kept. Here, inclusion levels rose from 30% to 50% (**Figure 8**B, lane 5). Importantly, analysis of the *MPDZ*$^{\Delta BS2}$ mutant has to be handled with caution. Cells transfected with the *MPDZ*$^{\Delta BS2}$ minigene had a high percentage of cell death and we only obtained little minigene mRNA from these cells. This might indicate that removing the downstream FUBP1 binding site results in a cytotoxic RNA species. A possible reason could be that deletion of the downstream binding site in *MPDZ*$^{\Delta BS2}$ affects exon 3 inclusion, as it acts as the upstream FUBP1 binding site of exon 3. When comparing exon 2 inclusion levels in minigenes with FUBP1 binding site deletions (*MPDZ*$^{\Delta BS1}$, *MPDZ*$^{\Delta BS2}$, and *MPDZ*$^{\Delta BS}$), we saw that *MPDZ*$^{\Delta BS1}$ and *MPDZ*$^{\Delta BS}$ are almost identical. This indicates that the downstream FUBP1 plays a subordinate role in inclusion levels compared to the upstream FUBP1 binding site. Overall, though the downstream binding site is not the driving force in exon inclusion, it still partakes in the splicing decision.

We established that FUBP1 binding and FUBP1 binding sites play a role in splicing decisions. Now, we also aim to understand whether the length of the exon changes the *MPDZ* minigene splicing pattern. In our study in chapter 4, shortening of both the upstream and the downstream introns did not substantially alter exon 2 inclusion levels (**Figure 8**B). Surprisingly, splicing of *MPDZ*$^{\Delta intron1}$

showed that deleting intronic sequences within the upstream intron changed splicing patterns (**Figure 8**B, lane 3). Shortening the downstream exon (*MPDZ*$^{\Delta intron2}$), however, did not change inclusion levels of exon 2 compared to the original *MPDZ* minigene (**Figure 8**B, lane 6). We hypothesize that not the intron length influences splicing, but there might be other regulatory elements in the upstream intron, such as intronic splicing enhancers and silencers, which can influence splicing decisions.

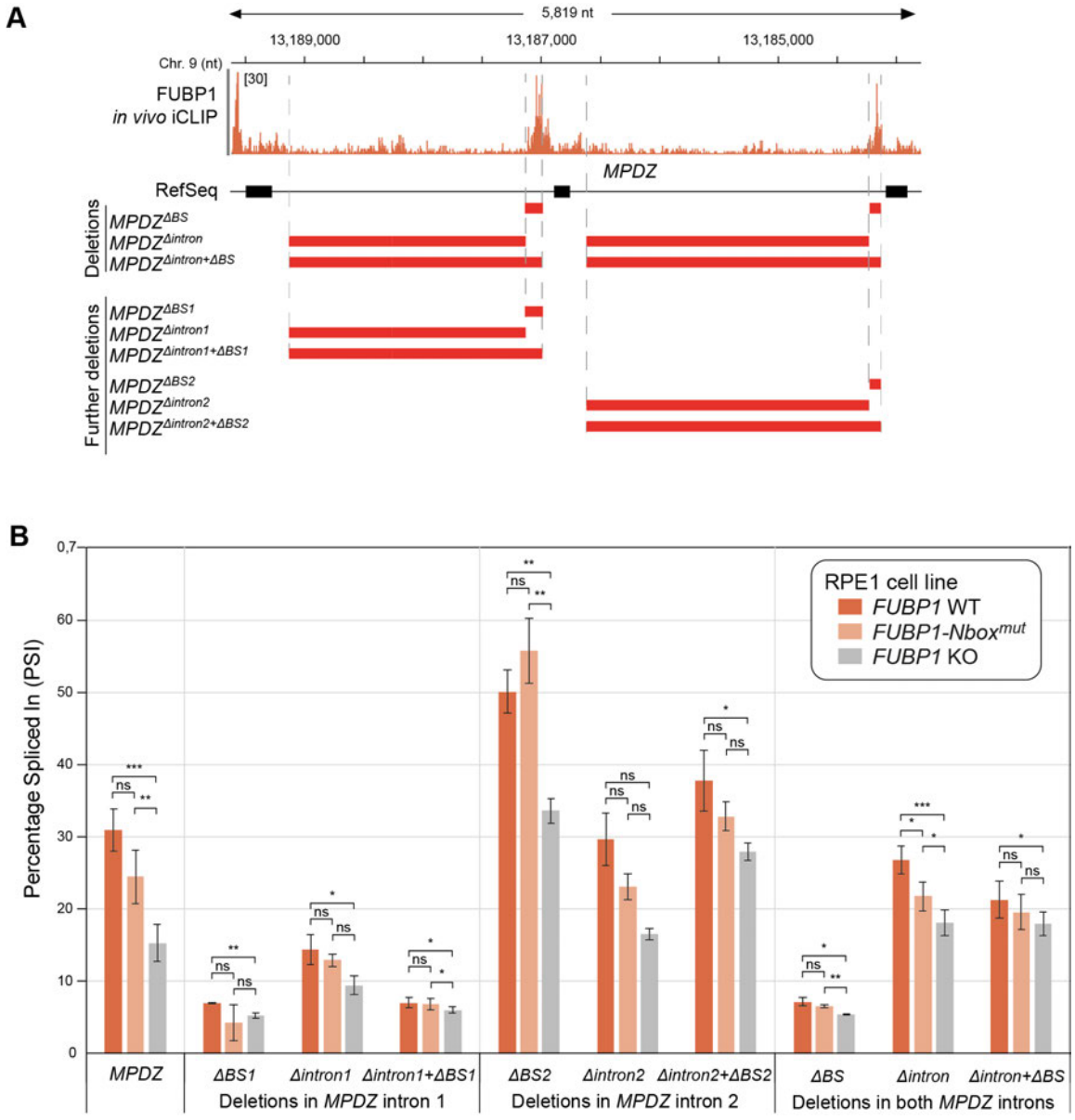


Figure 8: Minigene splicing assay with *MPDZ* minigene variants. (A) Description of all used minigene variants. The "Deletions" variants as well as their results were published in chapter 4. The "Further deletions" have been tested in the same experiments, but have not been described before. (B) Inclusion levels of alternative exon 2 using all described minigene variants. Minigenes were transfected and after RT-PCR, the percentage spliced in (PSI) was measured by capillary electrophoresis. For *MPDZ, MPDZ*$^{\Delta intron}$, *MPDZ*$^{\Delta intron+\Delta BS}$, n=6, for all other minigenes n=3. Significance determined by Student's t-test and corrected for multiple testing according to Benjamini and Hochberg. * p-value <0.05, ** p-value <0.01, *** p-value <0.001.

Taken together, our data shows that FUBP1 binding to the upstream binding site drives exon 2 inclusion. Though we see that FUBP1 binding to the downstream binding site is also important, it has a subordinate effect to the upstream binding site. The length of the intron itself might not play a role in exon inclusion, but rather the strength of the FUBP1 within these long introns, as explained in chapter 4. Whether this conclusion can be drawn on a transcriptome-wide level, however, needs to be further investigated.

6. Discussion

6.1. Overview

This work elaborates on two studies with the same goal: To unravel detailed splicing regulation in health, disease and evolution. First, we investigated how splicing is involved in the success and failure of CART-19 therapy of B-ALL patients. This further added detail to how CD19 can undergo epitope loss in order to evade the CART-19 therapy (Maude et al. 2016; Orlando et al. 2018). Second, we functionally characterized FUBP1 as a new splicing factor. In contrast to previous works, we could show that FUBP1 is a key splicing factor, adding another layer to its various functions in both DNA and RNA binding (Debaize and Troadec 2019). In these studies, we applied a multitude of techniques, ranging from molecular biology in form of minigene splicing assays over biochemical NMR-based approaches to mathematical modelling as well as high-throughput technologies. Ultimately, our findings highlight the interdependency of *cis*-regulatory sequences and the corresponding *trans*-acting factors. These communication networks can shape exon-intron architecture across the phylogenic tree; moreover, specific splicing landscapes in diseases like cancer are an ideal target for treatment (Frankiw, Baltimore, and Li 2019; Lynch and Richardson 2002).

It has previously been shown that FUBP1 has some implications in splicing regulation, however, its global role is poorly understood (Debaize and Troadec 2019; Hwang et al. 2018; Jacob et al. 2014; H. Li et al. 2013; Miro et al. 2015; J. Wang, Schultz, and Johnson 2018). We showed that FUBP1 is a core splicing factor that facilitates splicing of long introns. FUBP1 ubiquitously binds pre-mRNA to a similar extent as U2AF2 and SF3B1. Combining NMR- and CLIP-based methods, we discovered a new motif upstream of the branch point that is bound by FUBP1. Several protein interaction assays revealed that the N-terminal N-box of FUBP1 interacts with the RRM2 domain of U2AF2, and this interaction is vital for U2AF2 stabilization to the py-tract. We found that the C-terminal A/B-boxes are an animal-specific motif that are exclusive to the FUBP family. These A/B-boxes interact with proline-rich regions and facilitate the interaction of FUBP1 to SF1. Further, we uncovered that FUBP1 loosely interacts with components of the U1 snRNP like SNRPA or TIAL1.

Mechanistically, we showed that FUBP1 mediates splicing of long introns. FUBP1 interacted more strongly with long introns, as those harbor stronger FUBP1 binding motifs. When FUBP1 is depleted, exons flanked by long introns were prone to being skipped. Minigene splicing assay showed that this is a direct effect of FUBP1 binding to those introns and that without FUBP1 binding sites, it is indifferent whether FUBP1 is present or not. Interestingly, patients with low-grade glioma expressing loss-of-function *FUBP1* also showed decreased exon inclusion, validating our results (Seiler et al. 2018). Taken together, we hypothesize that FUBP1 engages in interaction with components of both the 3'ss and 5'ss to facilitate splice site bridging, which is a challenging task and especially important for long introns (Pai et al. 2017; Wachutka et al. 2019).

Our unpublished data of the *MPDZ* minigene assay revealed that the upstream intron predominantly regulates alternative exon inclusion levels. However, we only showed this in context of a single minigene. Further, we hypothesized that FUBP1 might influence splicing speed. Because splicing differs from long to short exons, their kinetic might also be different (De Conti, Baralle, and Buratti 2013; Pai et al. 2017; Wachutka et al. 2019). As a crucial splicing factor for long introns, FUBP1 could affect the splicing kinetic of long introns. Unexpectedly, examination of nascent pre-RNA did not show changes in splicing speed when FUBP1 was depleted. This might be due to the measurement of background noise, but could also indicate functions that have not been investigated yet. It will require further effort to fully understand the influence of FUBP1 on splicing.

In addition to how FUBP1 uses the splicing code to regulate splicing, we analyzed the splicing code on *CD19* mRNA to understand how B-ALL patients develop resistance to CART-19 therapy (Orlando et al. 2018). Using a high-throughput mutagenesis approach, we discovered that the usage of cryptic splice sites gives rise to approximately 100 previously unknown splice isoforms, most of which encode non-functional CD19 proteins. Furthermore, we identified RNA-binding proteins like PTBP1 and SF3B4 to regulate *CD19* mRNA splicing and loss of those lead to aberrant *CD19* splicing. Taken together, this study yields a detailed description of possible biomarkers to be used in B-ALL therapy, but it also highlights the importance of splicing in cancer and cancer treatment.

During my PhD, I primarily focused on unraveling the functional characteristics of FUBP1. In the following, I will discuss the findings in chapter 4 and 5 concerning the role of FUBP1 in splicing, evolution and disease, in light of the current research of the splicing field. I first evaluate the importance and novelty of our findings compared to current literature. Then, I discuss similarities in the known and novel interactions that FUBP1 maintains. Next, I put our findings in the different steps of splicing into context, starting at the definition of splice sites, moving to intron bridging and then the co-transcriptional splicing process. Throughout the discussion, I will not only focus on the different mechanisms in human, but also across evolution, especially in the model organisms *D. melanogaster* and yeast. I will end this discussion with a more global view on this thesis. I will look on our used *FUBP1* mutant cell lines, and finally, provide a link between our research and the rather philosophical Drift Barrier hypothesis.

6.2. FUBP1 is a novel core splicing factor

In our study, we discovered FUBP1 as a core splicing factor. Previously, it has been shown that FUBP1 is associated with the spliceosome (Rappsilber et al. 2002) and that it influences differential inclusion of single exons (Hwang et al. 2018; Jacob et al. 2014; H. Li et al. 2013; Miro et al. 2015; J. Wang, Schultz, and Johnson 2018). Contrary to previous studies, our work presents the first transcriptome-wide functional analysis of FUBP1. Our findings set FUBP1 into the company of

approximately 150 spliceosomal and splicing-associated proteins, and the number is rising constantly. For example, only this year, the proteins MEN1 and PRPF39 have been identified as regulators of alternative splicing (Espinosa et al. 2023; Jin et al. 2023). Additionally, further studies explored new splicing regulators in other model organisms like *A. thaliana* (MDF), *M. musculus* (NRDE2, CCDC174) and *D. melanogaster* (Tip60), to only name a few (Bhatnagar et al. 2023; Flemr et al. 2023; de Luxán-Hernández et al. 2023). Importantly, none of the mentioned splicing factors has been identified as core splicing factors. Even though literature provides an excellent overview of how the biochemical process of splicing works, we only begin to understand spliceosome assembly and splice site recognition (Nilsen and Graveley 2010).

In addition to establishing FUBP1 as a novel core splicing factor, we found the corresponding binding motif, a UUU+R stretch (R = A or G). Interestingly, this new motif is present transcriptome-wide upstream of the branch point and is especially pronounced in long introns. Concerning our observation that *FUBP1* is present in species with longer introns and that A/B-boxes are animal-specific, it will be interesting to see whether this *cis*-regulatory motif is also conserved in species that contain the *FUBP1* gene or A/B-boxes. For SF1, for example, we see such a conservation of *cis*-regulatory sequence together with the *trans*-acting element. In lower eukaryotes like yeast, the SF1 homolog BBP also recognizes the branch point. However, though yeast also expresses the U2AF2 homolog Mud2, the py-tract is not conserved. Mud2 and the U2AF1 homolog Msl5 only recognize the 3'ss (Wilkinson, Charenton, and Nagai 2020). Because we find evidence of both independent and co-evolved splicing elements, we cannot predict the conservation of the U-rich FUBP1 binding sites.

Why do we discover these splicing regulators and motifs only now, and continue to find new ones? Especially with FUBP1 being a core splicing factor, it is surprising that its role as an RBP is so poorly understood. There are two answers to this question. Naturally, advances in technologies are made faster and faster. For example, Sanger invented his sequencing method 1977 (Sanger, Nicklen, and Coulson 1977). Next generation sequencing was first published 20 years later in 1996, and only twelve more years later, in 2012 the next technological jump, the Nanopore sequencing, has been presented (Check Hayden 2012; Ronaghi et al. 1996). Using innovative tools like iCLIP, CRISPR/Cas9 assays coupled to RNA-seq, or mathematical modeling, provides a vast amount of data. All this is only possible due to the recent advances in biotechnology. Even more important than technology is the choice of the right model organism. Even though the knowledge about the conservation of splicing has been made in the early days of splicing research, most studies on splicing have been performed with yeast models (Meyer and Vilardell 2009). However, we know now that the underlying mechanisms in yeast and humans drastically changed over the course of evolution (Wilkinson, Charenton, and Nagai 2020). To this end, in order to understand splicing in a given organism, we have to study this specific species.

6.3. FUBP1 interaction interfaces potentially shape the functions of FUBP1

Using NMR- and BRET-based protein-protein-interaction assays, we demonstrated that FUBP1 maintains multivalent interactions with proteins on both the 3'ss and the 5'ss. In particular, our structural analyses pinpoint the interaction interfaces between FUBP1 and the 3' proteins SF1 and U2AF2. Furthermore, we compared the interaction of FUBP1 with U2AF2 to the interaction that FUBP1 maintains with FIR. Intriguingly, these two interactions are structurally similar. In both cases, the helical N-box of FUBP1 interacts with the RRM2 domain of the antagonist. An alignment of both RRM2 domains showed a high overlap between U2AF2 and FIR. Our iCLIP data revealed that the interaction of FUBP1 to U2AF2 is necessary for U2AF2 to be recruited to the py-tract. This FUBP1 dependency can also be found for the functionality of FIR (Debaize and Troadec 2019). Without interacting with FUBP1 first, FIR would not be capable of suppressing FUBP1-mediated transcription. We know the impact of FUBP1 on the role of U2AF2, but so far, we do not understand the influence of U2AF2 for the role of FUBP1. Therefore, it will be interesting to see how FUBP1-dependent splicing patterns change in transient depletion of U2AF2, e.g. by endogenous tagging of U2AF2 with a degron tag, and whether they might mimic FUBP1-FIR interaction characteristics.

FUBP1 contains A/B-boxes, a domain that has been found in *D. melanogaster*, but we could show that it is a FUBP-family specific domain in animals, including humans. In fruit flies, it was shown that A/B-boxes interact with proline-rich sequences (Ignjatovic et al. 2005; Labourier, Adams, and Rio 2001). This, however, remains to be proven in humans. Specifically, though we showed that FUBP1 loosely interacts with the U1 snRNP and U1-related proteins like SNRPA and TIAL1, it will be enlightening to investigate whether these interactions are maintained by the A/B-boxes of FUBP1 and the proline-rich sequences of U1 proteins. FUBP1 also contains proline-rich stretches, which could potentially interact with the numerous WW domains in proteins of the PRPF40 family, like PRPF40A and PRPF40B, but also TCERG1 and FBP21 (Ingham et al. 2005).

Not only protein-protein interaction interfaces can influence the role of FUBP1, also secondary structures on RNA can alter FUBP1 function. For example, a hairpin structure upstream of FUBP1-bound exonic splicing silencer has been shown to be critical for FUBP1-mediated skipping events (Li 2013). Long introns, which are generally sensitive to FUBP1 binding, have the potential to form complex secondary structures. However, whether there are secondary structures that aid FUBP1 as a general splicing factor remains to be shown.

6.4. Involvement of FUBP1 exon definition

The splicing process commences with the recognition and definition of the exon and intron boundaries. This means that the U1 snRNP and the U2 snRNP bind to the 5'ss and 3'ss, respectively, and recognize the exon-intron boundaries (De Conti, Baralle, and Buratti 2013). The U1 and U2

snRNP then interact with each other to define the splicing units. For short introns, this interaction is facilitated over the intron, and therefore the process is called intron definition. When long introns are present, however, bridging the distance between the two opposing splice sites is a difficult task. Therefore, the initial contact between U1 and U2 components is first facilitated over the exon and subsequently, over the intron. This mode of action is called exon definition (Susan M Berget 1995; De Conti, Baralle, and Buratti 2013; Keren, Lev-Maor, and Ast 2010). Coupled to exon definition is an increase in splicing speed. In *D. melanogaster*, a species containing long introns, it was found that intron-defined long introns are spliced slower than intron-defined short introns. Intriguingly, exon-defined long introns were spliced faster than short introns (Pai et al. 2017). Indeed, this has also been shown to be true for human splicing (Wachutka et al. 2019). Further, insufficient exon definition leads to exon skipping (Tammer et al. 2022).

Our data clearly revealed that FUBP1 is crucial for splicing of long introns. FUBP1 binding sites are more pronounced in long introns and therefore, FUBP1 binds more prevalently to them. In absence of FUBP1, exons flanked by long introns are aberrantly skipped. This is true for both an *FUBP1* KO cell line and *FUBP1* loss-of-function cancer patients. Our minigene analysis showed that this is a direct effect of FUBP1 binding to the flanking long introns. We hypothesize that FUBP1 aids in exon definition and that without FUBP1, exon cannot be properly defined, which leads to exon skipping. As exon-defined long introns are spliced faster than intron-defined intron, loss of FUBP1 would lead to a decrease in splicing speed. To investigate this hypothesis, we performed nascent RNA-seq. We labeled *FUBP1*-depleted cells and WT cells for 1h with 4sU and subsequently isolated the labeled RNA to obtain RNA that has been transcribed within the hour of labeling. However, we could not see any correlation between loss of FUBP1 and splicing speed, but most likely, we measured background noise, as splicing happens in order of minutes, not hours (Alpert, Herzel, and Neugebauer 2017). A more fine-tuned nascent RNA-seq like TT-seq with shorter labeling times might shed further light onto the sway of FUBP1 on splicing kinetics and exon definition (Schwalb et al. 2016).

6.5. The FUBP1 interactions add a new layer of splice site bridging

Through protein-protein-interaction assays, we could identify several targets of FUBP1. As explained above, FUBP1 interacts with proteins of the 3'ss like U2AF2 and SF1, but it also interacts with U1 snRNP proteins like SNRPA and TIAL1. Additionally, we know that FUBP1 acts on long introns, where the splice sites are very far apart. Therefore, we hypothesize that after exon definition, FUBP1 facilitates bridging of long introns in order to bring the 5'ss and the 3'ss into close proximity. These interactions between FUBP1 and the U1 components are rather weak. It would be astonishing if these interactions could maintain splice site connection over the distance of several thousand nucleotides to connect two exons. Naturally, FUBP1 is not the only protein facilitating intron

bridging. For example it has been shown that both human PRPF40A and PRPF40B can interact with 3'ss components like U2AF2 and SF1 (Becerra et al. 2015; Lin, Lu, and Tarn 2004). This also has been shown for the yeast homologs, where Prp40 interacts with Mud2 and BBP, respectively (Abovich and Rosbash 1997). On the other hand, it has been shown that Prp40 also interacts with components of the U1 snRNP, like LUC7L and RBM25 (Abovich and Rosbash 1997; X. Li et al. 2019; Plaschka et al. 2018). SF1 however, detaches from the spliceosome quite early, leaving room for further interactions between the U1 snRNP and the 3'ss. We believe that FUBP1 takes over the role of SF1 in maintaining the contacts to the other splice site by interacting with its A/B-boxes with the multiple proline-rich sequences in U1 snRNP proteins and associated proteins. In addition, there are multiple other interaction interfaces between the U1 and U2 snRNP. For example, it has been shown that the U1 snRNA interacts with SF3A1 and RNA helicase DDX39B (Martelly et al. 2021; Sharma et al. 2014). In the end, the splice site bridging complex can be viewed as an accumulation of loosely interacting components with multivalent, ever-changing, interaction partners.

6.6. Solutions to excising long introns across species

Not only *trans*-acting factors facilitate intron bridging, but also sequences within the pre-mRNA itself help to bring two exons into close proximity. Complementary Alu sequences within an intron can base-pair and thereby bridge long intronic sequences. When complementary Alus are situated at opposing sites of an intron, they positively influence inclusion levels of the flanking exons (Lev-Maor et al. 2008). Alu sequences are inherently U-rich and could harbor FUBP1 binding motifs (Batzer and Deininger 2002). Whether FUBP1 binds to intronic Alu sequences, however, remains to be investigated. Nevertheless, it is easy to picture FUBP1 binding to Alu sequences and thereby exerting its role in intron bridging, especially because both Alus and the A/B-boxes are a trait of higher eukaryotes (Batzer and Deininger 2002; Kriegs et al. 2007). A simple way to investigate the connection between FUBP1 and Alus can be provided by the *MPDZ* minigene. *MPDZ* also harbors complementary Alu sequences, which could be artificially excluded from the minigene. Subsequent transfection of the minigene to *FUBP1* WT and KO cells followed by RT-PCR could offer valuable insights into Alu-driven FUBP1 functions. In addition, searching existing iCLIP data for FUBP1-bound Alus might shed further light on this topic. How conserved such a mechanism could be, however, seems debatable, as the human genome harbors a substantial quantity of large and small interspersed nucleotide elements (LINE/SINE) (33%), which is not the case for other species with long introns (International Human Genome Consortium 2001). For example, *D. melanogaster* expresses the FUBP1 homolog Psi, which also has been linked to splicing due to its interaction with the U1 protein SNRNP70 (Ignjatovic et al. 2005; Labourier, Adams, and Rio 2001). However, the fruit fly genome consists only to 0.7% of LINE/SINE sequences (International Human Genome Consortium 2001).

Fruit flies possess another approach on bridging long introns. *D. melanogaster* also displays a large range of long introns, but instead of excising the intron in one piece, it cuts out the intron in short segments. This process is called recursive splicing. Until recently, there were only a few recursive splicing events known in humans. However, this year, Hoppe and colleagues used an impressively elegant bioinformatic analysis on RNA-seq data to find recursive splicing events in humans. Their study reveals 100 hitherto unknown recursive splicing events, demonstrating how underestimated human recursive splicing was so far (Hoppe, Udy, and Bradley 2023). By re-analyzing existing RNA-seq data or nascent RNA-seq data of RPE1 WT and *FUBP1* KO cell lines, we could approach a more comprehensive picture of the mechanisms with which FUBP1 impacts splicing.

6.7. Transcription influences splicing dynamics and splicing outcome

During nascent RNA-sequencing, we recovered more nascent RNA in WT samples compared to *FUBP1*-depleted samples. *FUBP1*-depleted cells also grew slower, which is in line with literature stating that FUBP1 regulates cell growth (Debaize and Troadec 2019). This could indicate that in *FUBP1* KO, the transcription rate is much lower. FUBP1 interacts with TFIIH and thereby influences PolII release, indicating a role in transcription rate (Quinn 2017). Even though this might not seem directly related to splicing, it has been shown that transcription strongly affects splicing decisions and an increasing number of studies concerning co-transcriptional splicing enrich our understanding of the link between these two mechanisms (Herzel et al. 2017). For example, we now know that U1 snRNP interacts with the C-terminal domain (CTD) of PolII during transcription and thereby increases elongation rate for long genes (Mimoso and Adelman 2023; Zhang et al. 2021). In addition, the nascent 5'ss also remains tethered to PolII until the end of the intron is transcribed (Leader et al. 2021).

Whether FUBP1 indeed influences transcription speed, still needs to be investigated. To this end, we would require a measure for PolII speed in WT and *FUBP1* KO conditions, for example ChIP-seq or GRO-seq (Jonkers and Lis 2015). This would offer the possibility to study the role of FUBP1in transcription and in combination with our *FUBP1* KO RNA-seq and a potential nascent RNA-seq using TT-seq, could result in new insights into the details of co-transcriptional splicing. A more simple and elegant strategy to account for transcription rate and splicing speed simultaneously, has been established by Singh and Padgett (J. Singh and Padgett 2009). Here, the reversible transcription inhibitor 5,6-Dichlorobenzimidazole 1-b-D-ribofuranoside (DRB) was used to rapidly switch transcription off and on. In conjunction with RT-qPCR of exon-intron boundaries, they could not only measure transcription speed, but also splicing velocity. This method has also been employed in a transcriptome-wide manner, combining DRB treatment with a 4sU-labeling time course (Fuchs et al. 2014).

Another factor that heavily influences co-transcriptional splicing is histone occupancy. Exons are generally more likely to be packaged into nucleosomes. Constitutive exons are bound by histones carrying H3K36me3 modification, slowing down PolII, and enabling the 5'ss to be tethered to the CTD of PolII. When the methylation is missing, e.g. for tissue-specific alternative splicing events, PolII passes the exon rapidly, leading to a failed tethering of the 5'ss to the CTD, which in turn leads to a exon skipping event (reviewed in Nilsen and Graveley 2010). Our results clearly show that FUBP1 binds RNA upstream of the 3'ss, but it has not yet been shown whether FUBP1 also binds DNA at the same position, which would correspond to a position directly upstream of a nucleosome. It would also be worthwhile to investigate the interactions between FUBP1, histones and histone methylases. This might unravel further mechanistic insights into the role of FUBP1 in splicing decisions and co-transcriptional splicing.

6.8. Evaluation of the *FUBP1* KO as a model to study FUBP1-specific splicing regulation

Though we were very delighted to have achieved the elegant *FUBP1* KO and *FUBP1 N-box*^{mut} cell lines, this model system came with challenges. As we have shown, FUBP1 is crucial for splicing of long introns. In addition, FUBP1 regulates several molecular and cellular processes like transcription and translation, as well as cell growth and proliferation (Debaize and Troadec 2019). Moreover, it has repeatedly been shown that a global *FUBP1* KO in mice is lethal starting from embryonic day 10.5 (Hwang et al. 2018; Zhou et al. 2016). Therefore, a *FUBP1* KO in cells is expected to have severe effects on the cellular viability. However, other than a decreased velocity in cell growth, our *FUBP1* KO cell lines appeared healthy. This indicates that there might be compensation mechanism within the cells that we were not able to identify. Especially the homolog KHSRP might rescue *FUBP1* KO-mediated phenotypes, as it has been observed that a viable *FUBP1* KO cell line cannot be sustained when *KHSRP* is transiently knocked down simultaneously (Ying Zheng et al. 2020). Both FUBP1 and KHSRP can regulate MYC expression, and they have high sequence similarity (Debaize and Troadec 2019; Ying Zheng et al. 2020). KHSRP is capable of binding RNA within the nucleus. More specifically, it can interact with FUSE sequences and AU-rich elements, which are key features of FUBP1 binding motifs (Gherzi et al. 2010). Additionally, KHSRP has been linked to exon inclusion by binding to an intronic splicing enhancer (Min et al. 1997). Though we could not observe upregulation of KHSRP in RNA-seq of *FUBP1* KO cells, we cannot exclude the possibility that KHSRP partakes in FUBP1-specific roles.

For a number of experiments, we used a minigene to exemplify the effect of FUBP1 in splicing. We used a sequence of the *MPDZ* gene, harboring chr9:13,183,353-13,189,041. This sequence contains two long introns separating three exons, of which the second is alternatively spliced. With the minigene, we could show that the inclusion of the alternative exon depends on FUBP1 binding to its

motif upstream to the 3'ss. In addition, we were able to rescue the *FUBP1* KO by overexpression of exogenous FUBP1. To a lesser extent, this was possible with truncated FUBP1 constructs, indicating that all interactions facilitated by FUBP1 are important to maintain its function as a splicing factor. Though these are striking results, we could only show this effect on the *MPDZ* minigene, but were unable to replicate the results with endogenous *MPDZ* or other FUBP1 targets. Therefore, we conclude that the *FUBP1* KO cell line comprised mechanisms to shortcut FUBP1, which could influence the results obtained with the *FUBP1* KO cell line. Therefore, [redacted], a member of both the [redacted] and [redacted] group, will continue this study by establishing an *FUBP1* KO inserting a degron tag to endogenous FUBP1. Thereby, FUBP1 can rapidly be degraded, creating a transient, yet effective knockdown, combining the efficiency of a stable knockout and the speed of a siRNA-mediated knockdown.

6.9. Splicing evolution and the Drift Barrier hypothesis

Michael Lynch became famous by claiming that the better a trait performs in a specific environment, the smaller the fitness advantages become by additional beneficial mutations. Meaning, that when a trait performs perfectly in a certain environment, additional mutations will not enhance the performance of that perfect trait. Ultimately, natural selection will hit a hypothetical ceiling that was termed the Drift Barrier (Lynch 2010). If there is no added benefit, all beneficial mutations only spread through genetic drift, which means it only fixes in a population through chance. However, the Drift-Barrier hypothesis is a genetic and evolutionary assumption that is difficult to prove.

It is widely understood that splicing evolved with the last eukaryotic common ancestor (LECA). However, splicing strongly differs from human to yeast, ranging from intron length over exon and intron definition models to splicing complexity. In light of evolution, one must wonder whether the benefit of long introns, together with the added complexity with enhancers, silencers and other regulators, actually outweighs the immense cost that come with such a massive splicing machinery. It might just be that in higher eukaryotes, splicing hit the drift barrier, unable to loose long introns and instead evolved to use the space for regulation by constructive neutral evolution, making splicing ever more complex. Therefore, FUBP1 might be the evidence that animals are stagnant in splicing evolution at the Drift Barrier, unable to get rid of intron length and instead, by constructive neutral evolution, emerged to bridge longer introns, rather than making bridging easier.

7. Conclusion and Outlook

In this study, we discovered FUBP1 as a novel core splicing component. However, our study opens further questions. First, even though we know that the presence of FUBP1 is necessary for exon inclusion, we do not know which mechanisms are employed to facilitate inclusion. We hypothesize that through the interactions with components of the 3'ss and 5'ss, FUBP1 can aid in exon definition and subsequent intron bridging and thereby bring the active site of the spliceosome to the right position. However, to date, we lack proof that these interactions are crucial for FUBP1 functionality. In addition, we still do not understand the implications of FUBP1 in splicing kinetics. Performing different nascent RNA-sequencings with shorter time points will be crucial for determining whether FUBP1 plays a role in co-transcriptional splicing and in splicing speed.

On an evolutionary level, we could show that splicing regulation by FUBP1 is associated with many features specific to higher eukaryotes, e.g. long introns and A/B-boxes. This gives a glimpse of how far splicing has evolved from the last common eukaryotic ancestor. To understand the differences in splicing systems from yeast to human, more research is needed. This can involve understanding the conservation of FUBP1 as a splicing protein in other species, but also to further study recursive splicing in humans.

Ultimately, these studies help understand splicing on a more profound level. Aberrant splicing is a well-known hallmark of cancer, and understanding splicing is a crucial part of cancer treatment. We showed that CD19 can lose its epitope through mis-splicing, leading to relapse of CART-19 therapy. FUBP1 misregulation is a common feature in many cancers; and in low-grade glioma, we saw a FUBP1-specific splicing impairment. These studies can help advance treatments for cancer and other diseases like spinal muscular atrophy. Therefore, this thesis adds important knowledge to medicine; however, we also show that there is more research needed before we fully understand the splicing code.

8. References

Abovich, Nadja, and Michael Rosbash. 1997. “Cross-Intron Bridging Interactions in the Yeast Commitment Complex Are Conserved in Mammals.” *Cell* 89(3): 403–12.

Alpert, Tara, Lydia Herzel, and Karla M Neugebauer. 2017. “Perfect Timing: Splicing and Transcription Rates in Living Cells.” *WIREs RNA* 8: 1401. https://wires.onlinelibrary.wiley.com/doi/10.1002/wrna.1401 (October 31, 2022).

Alsafadi, Samar et al. 2016. “Cancer-Associated SF3B1 Mutations Affect Alternative Splicing by Promoting Alternative Branchpoint Usage.” *Nature communications* 7: 10615.

Alt, Frederick W et al. 1980. “Synthesis of Secreted and Membrane-Bound Immunoglobulin Mu Heavy Chains Is Directed by MRNAs That Differ at Their 3′ Ends.” *Cell* 20(2): 293–301. https://doi.org/10.1016/0092-8674(80)90615-7.

Amit, Maayan et al. 2012. “Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition.” *Cell Reports* 1(5): 543–56. https://www.sciencedirect.com/science/article/pii/S2211124712000988.

Asnani, Mukta et al. 2020. “Retention of CD19 Intron 2 Contributes to CART-19 Resistance in Leukemias with Subclonal Frameshift Mutations in CD19.” *Leukemia* 34(4): 1202–7.

Atanassov, Boyko S, and Sharon Y R Dent. 2011. “USP22 Regulates Cell Proliferation by Deubiquitinating the Transcriptional Regulator FBP1.” *EMBO reports* 12(9): 924–30. https://doi.org/10.1038/embor.2011.140.

Avigan, M I, B Strober, and D Levens. 1990. “A Far Upstream Element Stimulates C-Myc Expression in Undifferentiated Leukemia Cells.” *The Journal of biological chemistry* 265(30): 18538–45.

Bagashev, Asen et al. 2018. “CD19 Alterations Emerging after CD19-Directed Immunotherapy Cause Retention of the Misfolded Protein in the Endoplasmic Reticulum.” *Molecular and Cellular Biology* 38(21): 1–17.

Baralle, Francisco E, and Jimena Giudice. 2017. “Alternative Splicing as a Regulator of Development and Tissue Identity.” *Nature Reviews Molecular Cell Biology* 18(7): 437–51. https://doi.org/10.1038/nrm.2017.27.

Batzer, Mark A., and Prescott L. Deininger. 2002. “Alu Repeats and Human Genomic Diversity.” *Nature Reviews Genetics* 3(5): 370–79.

Becerra, Soraya, Marta Montes, Cristina Hernández-Munain, and Carlos Suñe. 2015. “Prp40 Pre-MRNA Processing Factor 40 Homolog B (PRPF40B) Associates with SF1 and U2AF65and Modulates Alternative Pre-MRNA Splicing in Vivo.” *Rna* 21(3): 438–57.

Ben-Dov, Claudia, Britta Hartmann, Josefin Lundgren, and Juan Valcárcel. 2008. “Genome-Wide Analysis of Alternative Pre-MRNA Splicing * .” *Journal of Biological Chemistry* 283(3): 1229–33. https://doi.org/10.1074/jbc.R700033200.

Bentley, David L. 2014. “Coupling MRNA Processing with Transcription in Time and Space.” *Nature Reviews Genetics* 15(3): 163–75. https://doi.org/10.1038/nrg3662.

Berget, S. M., C. Moore, and P. A. Sharp. 1977. “Spliced Segments at the 5’ Terminus of Adenovirus 2 Late MRNA.” *Proceedings of the National Academy of Sciences of the United States of America* 74(8): 3171–75.

Berget, Susan M. 1995. “Exon Recognition in Vertebrate Splicing (∗).” *Journal of Biological Chemistry* 270(6): 2411–14. https://doi.org/10.1074/jbc.270.6.2411.

Bergsma, Atze J et al. 2018. “Chapter Three - Alternative Splicing in Genetic Diseases: Improved

Diagnosis and Novel Treatment Options." In *Transcriptional Gene Regulation in Health and Disease*, ed. Friedemann B T - International Review of Cell and Molecular Biology Loos. Academic Press, 85–141. https://www.sciencedirect.com/science/article/pii/S1937644817300837.

Bertrand, F E, C Vogtenhuber, N Shah, and T W LeBien. 2001. "Pro-B-Cell to Pre-B-Cell Development in B-Lineage Acute Lymphoblastic Leukemia Expressing the MLL/AF4 Fusion Protein." *Blood* 98(12): 3398–3405.

Bhatnagar, Akanksha et al. 2023. "Tip60's Novel RNA-Binding Function Modulates Alternative Splicing of Pre-MRNA Targets Implicated in Alzheimer's Disease." *The Journal of Neuroscience* 43(13): 2398 LP – 2423. http://www.jneurosci.org/content/43/13/2398.abstract.

Blencowe, Benjamin J. 2006. "Alternative Splicing: New Insights from Global Analyses." *Cell* 126(1): 37–47. https://doi.org/10.1016/j.cell.2006.06.023.

Briata, Paola et al. 2016. "Diverse Roles of the Nucleic Acid-Binding Protein KHSRP in Cell Differentiation and Disease." *WIREs RNA* 7(2): 227–40. https://doi.org/10.1002/wrna.1327.

Caizzi, Livia et al. 2021. "Efficient RNA Polymerase II Pause Release Requires U2 SnRNP Function." *Molecular Cell* 81(9): 1920-1934.e9.

Check Hayden, Erika. 2012. "Nanopore Genome Sequencer Makes Its Debut." *Nature*. https://doi.org/10.1038/nature.2012.10051.

Chow, Louise T, Richard E Gelinas, Thomas R Broker, and Richard J Roberts. 1977. "An Amazing Sequence Arrangement at the 5′ Ends of Adenovirus 2 Messenger RNA." *Cell* 12(1): 1–8. https://doi.org/10.1016/0092-8674(77)90180-5.

Chung, Hye-Jung et al. 2006. "FBPs Are Calibrated Molecular Tools To Adjust Gene Expression." *Molecular and Cellular Biology* 26(17): 6584–97. https://doi.org/10.1128/MCB.00754-06.

De Conti, Laura, Marco Baralle, and Emanuele Buratti. 2013. "Exon and Intron Definition in Pre-MRNA Splicing." *Wiley Interdisciplinary Reviews: RNA* 4(1): 49–60.

Corioni, Margherita et al. 2011. "Analysis of in Situ Pre-MRNA Targets of Human Splicing Factor SF1 Reveals a Function in Alternative Splicing." *Nucleic acids research* 39(5): 1868–79.

Danecek, Petr et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10(2).

Davis-Smyth, Terri et al. 1996. "The Far Upstream Element-Binding Proteins Comprise an Ancient Family of Single-Strand DNA-Binding Transactivators*." *Journal of Biological Chemistry* 271(49): 31679–87. https://www.sciencedirect.com/science/article/pii/S0021925819790425.

Debaize, Lydie, and Marie Bérengère Troadec. 2019. "The Master Regulator FUBP1: Its Emerging Role in Normal Cell Function and Malignant Development." *Cellular and Molecular Life Sciences* 76(2): 259–81. https://doi.org/10.1007/s00018-018-2933-6.

Dlamini, Zodwa, Fortunate Mokoena, and Rodney Hull. 2017. "Abnormalities in Alternative Splicing in Diabetes: Therapeutic Targets." *Journal of Molecular Endocrinology* 59(2): R93–107. https://jme.bioscientifica.com/view/journals/jme/59/2/JME-17-0049.xml.

Dobin, Alexander et al. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics (Oxford, England)* 29(1): 15–21.

DT, Braddock et al. 2002. "Structure and Dynamics of KH Domains from FBP Bound to Single-Stranded DNA." *Nature* 415: 1051–56. http://www.wormbase.org/db/misc/paper?name=WBPaper00013019.

Duffy, Michael J, Shane O'Grady, Minhong Tang, and John Crown. 2021. "MYC as a Target for Cancer Treatment." *Cancer Treatment Reviews* 94: 102154.

https://www.sciencedirect.com/science/article/pii/S0305737221000025.

Duncan, Robert et al. 1994. “A Sequence-Specific, Single-Strand Binding Protein Activates the Far Upstream Element of c-Myc and Defines a New DNA-Binding Motif.” *Genes and Development* 8(4): 465–80.

Dvorak, Pavel, Vojtech Hanicinec, and Pavel Soucek. 2023. “The Position of the Longest Intron Is Related to Biological Functions in Some Human Genes .” *Frontiers in Genetics* 13. https://www.frontiersin.org/articles/10.3389/fgene.2022.1085139.

Early, P et al. 1980. “Two MRNAs Can Be Produced from a Single Immunoglobulin μ Gene by Alternative RNA Processing Pathways.” *Cell* 20(2): 313–19. https://doi.org/10.1016/0092-8674(80)90617-0.

Espinosa, Sara et al. 2023. “Human PRPF39 Is an Alternative Splicing Factor Recruiting U1 SnRNP to Weak 5’ Splice Sites.” *RNA* (29): 97–110.

Fernández-Nogales, Marta et al. 2016. “Faulty Splicing and Cytoskeleton Abnormalities in Huntington’s Disease.” *Brain Pathology* 26(6): 772–78. https://doi.org/10.1111/bpa.12430.

Flemr, Matyas et al. 2023. “Mouse Nuclear RNAi-Defective 2 Promotes Splicing of Weak 5’ Splice Sites.” *RNA (New York, N.Y.)* 29(8): 1140–65.

Fox-Walsh, Kristi L et al. 2005. 8 *The Architecture of Pre-MRNAs Affects Mechanisms of Splice-Site Pairing*. www.pnas.orgcgidoi10.1073pnas.0508489102 (February 15, 2021).

Frankish, Adam et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic acids research* 47(D1): D766–73.

Frankiw, Luke, David Baltimore, and Guideng Li. 2019. “Alternative MRNA Splicing in Cancer Immunotherapy.” *Nature Reviews Immunology* 19(11): 675–87. http://dx.doi.org/10.1038/s41577-019-0195-7.

Fuchs, Gilad et al. 2014. “4sUDRB-Seq: Measuring Genomewide Transcriptional Elongation Rates and Initiation Frequencies within Cells.” *Genome Biology* 15(5): R69. https://doi.org/10.1186/gb-2014-15-5-r69.

Gallego-Paez, L M et al. 2017. “Alternative Splicing: The Pledge, the Turn, and the Prestige.” *Human Genetics* 136(9): 1015–42. https://doi.org/10.1007/s00439-017-1790-y.

Gao, Kaiping, Akio Masuda, Tohru Matsuura, and Kinji Ohno. 2008. “Human Branch Point Consensus Sequence Is YUnAy.” *Nucleic acids research* 36(7): 2257–67.

Gartel, Andrei L, Michael S Serfas, and Angela L Tyner. 1996. “P21—Negative Regulator of the Cell Cycle.” *Proceedings of the Society for Experimental Biology and Medicine* 213(2): 138–49. https://journals.sagepub.com/doi/abs/10.3181/00379727-213-44046.

Gau, Bing-Huang, Tsung-Ming Chen, Yu-Heng J Shih, and H Sunny Sun. 2011. “FUBP3 Interacts with FGF9 3′ Microsatellite and Positively Regulates FGF9 Translation.” *Nucleic Acids Research* 39(9): 3582–93. https://doi.org/10.1093/nar/gkq1295.

Gerstberger, Stefanie, Markus Hafner, and Thomas Tuschl. 2014. “A Census of Human RNA-Binding Proteins.” *Nature Reviews Genetics* 15(12): 829–45.

Gherzi, Roberto et al. 2010. “The Role of KSRP in MRNA Decay and MicroRNA Precursor Maturation.” *Wiley Interdisciplinary Reviews: RNA* 1(2): 230–39.

Gilbert, Walter. 1978. “Why Genes in Pieces?” *Nature* 271(5645): 501. https://doi.org/10.1038/271501a0.

Gressel, Saskia, Katja Lidschreiber, and Patrick Cramer. 2019. “Transient Transcriptome

Sequencing : Experimental Protocol to Monitor Genome-Wide RNA Synthesis Including Enhancer Transcription." *Protocols.Io*: 1–20.

He, Liusheng, Achim Weber, and David Levens. 2000. "Nuclear Targeting Determinants of the Far Upstream Element Binding Protein, a c-Myc Transcription Factor." *Nucleic Acids Research* 28(22): 4558–65. https://doi.org/10.1093/nar/28.22.4558.

Herzel, Lydia, Diana S.M. Ottoz, Tara Alpert, and Karla M. Neugebauer. 2017. "Splicing and Transcription Touch Base: Co-Transcriptional Spliceosome Assembly and Function." *Nature Reviews Molecular Cell Biology* 18(10): 637–50. http://dx.doi.org/10.1038/nrm.2017.63.

Hoppe, Emma R., Dylan B. Udy, and Robert K. Bradley. 2023. "Recursive Splicing Discovery Using Lariats in Total RNA Sequencing." *Life Science Alliance* 6(7): 1–13.

Hwang, Inah et al. 2018. "Far Upstream Element-Binding Protein 1 Regulates LSD1 Alternative Splicing to Promote Terminal Differentiation of Neural Progenitors." *Stem Cell Reports* 10(4): 1208–21. https://doi.org/10.1016/j.stemcr.2018.02.013.

Ignjatovic, Tijana et al. 2005. "Structural Basis of the Interaction between P-Element Somatic Inhibitor and U1-70k Essential for the Alternative Splicing of P-Element Transposase." *Journal of Molecular Biology* 351(1): 52–65.

Ingham, Robert J et al. 2005. "WW Domains Provide a Platform for the Assembly of Multiprotein Networks." *Molecular and cellular biology* 25(16): 7092–7106.

International Human Genome Consortium. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 412(6846): 565–66.

Irwin, Nina, Veerle Baekelandt, Luda Goritchenko, and Larry I Benowitz. 1997. "Identification of Two Proteins That Bind to a Pyrimidine-Rich Sequence in the 3′-Untranslated Region of GAP-43 MRNA." *Nucleic Acids Research* 25(6): 1281–88. https://doi.org/10.1093/nar/25.6.1281.

Jabbour, Elias, Susan O'Brien, Marina Konopleva, and Hagop Kantarjian. 2015. "New Insights into the Pathophysiology and Therapy of Adult Acute Lymphoblastic Leukemia." *Cancer* 121(15): 2517–28.

Jacob, Aishwarya G. et al. 2014. "The Splicing Factor FUBP1 Is Required for the Efficient Splicing of Oncogene MDM2 Pre-MRNA." *Journal of Biological Chemistry* 289(25): 17350–64.

Jang, M et al. 2009. "Far Upstream Element-Binding Protein-1, a Novel Caspase Substrate, Acts as a Cross-Talker between Apoptosis and the c-Myc Oncogene." *Oncogene* 28(12): 1529–36. https://doi.org/10.1038/onc.2009.11.

Jin, Bangming et al. 2023. "MEN1 Is a Regulator of Alternative Splicing and Prevents R-Loop-Induced Genome Instability through Suppression of RNA Polymerase II Elongation." *Nucleic acids research* 51(15): 7951–71.

Jo, Bong-Seok, and Sun Shim Choi. 2015. "Introns: The Functional Benefits of Introns in Genomes." *Genomics Inform* 13(4): 112–18. https://doi.org/10.5808/GI.2015.13.4.112.

Jonkers, Iris, and John T Lis. 2015. "Getting up to Speed with Transcription Elongation by RNA Polymerase II." *Nature Reviews Molecular Cell Biology* 16(3): 167–77. https://doi.org/10.1038/nrm3953.

Kahles, André et al. 2018. "Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients." *Cancer Cell* 34(2): 211-224.e6.

Kastner, Berthold, Cindy L. Will, Holger Stark, and Reinhard Lührmann. 2019. "Structural Insights into Nuclear Pre-MRNA Splicing in Higher Eukaryotes." *Cold Spring Harbor Perspectives in*

Biology 11(11).

Keren, Hadas, Galit Lev-Maor, and Gil Ast. 2010. "Alternative Splicing and Evolution: Diversification, Exon Definition and Function." *Nature Reviews Genetics* 11(5): 345–55.

Klenerman, Paul, Vincenzo Cerundolo, and P Rod Dunbar. 2002. "Tracking T Cells with Tetramers: New Tales from New Tools." *Nature reviews. Immunology* 2(4): 263–72. http://dx.doi.org/10.1038/nri777 (January 24, 2016).

Kriegs, Jan Ole et al. 2007. "Evolutionary History of 7SL RNA-Derived SINEs in Supraprimates." *Trends in Genetics* 23(4): 158–61. https://www.sciencedirect.com/science/article/pii/S0168952507000376.

Labourier, Emmanuel, Melissa D. Adams, and Donald C. Rio. 2001. "Modulation of P-Element Pre-MRNA Splicing by a Direct Interaction between PSI and U1 SnRNP 70K Protein." *Molecular Cell* 8(2): 363–73.

Leader, Yodfat et al. 2021. "The Upstream 5′ Splice Site Remains Associated to the Transcription Machinery during Intron Synthesis." *Nature Communications* 12(1): 4545. https://doi.org/10.1038/s41467-021-24774-6.

Lev-Maor, Galit et al. 2008. "Intronic Alus Influence Alternative Splicing." *PLoS genetics* 4(9): e1000204.

Li, Huang et al. 2013. "Far Upstream Element-Binding Protein 1 and Rna Secondary Structure Both Mediate Second-Step Splicing Repression." *Proceedings of the National Academy of Sciences of the United States of America* 110(29): 2687–95.

Li, Xueni et al. 2019. "A Unified Mechanism for Intron and Exon Definition and Back-Splicing." *Nature* 573(7774): 375–80.

Lin, Kai-Ti, Ruei-Min Lu, and Woan-Yuh Tarn. 2004. "The WW Domain-Containing Proteins Interact with the Early Spliceosome and Participate in Pre-MRNA Splicing In Vivo." *Molecular and Cellular Biology* 24(20): 9176–85. https://doi.org/10.1128/MCB.24.20.9176-9185.2004.

de Luxán-Hernández, Cloe et al. 2023. "MDF Is a Conserved Splicing Factor and Modulates Cell Division and Stress Response in Arabidopsis." *Life science alliance* 6(1).

Lynch, Michael. 2010. "Evolution of the Mutation Rate." *Trends in Genetics* 26(8): 345–52. https://www.sciencedirect.com/science/article/pii/S0168952510001034.

Lynch, Michael, and Aaron O. Richardson. 2002. "The Evolution of Spliceosomal Introns." *Current Opinion in Genetics & Development* 12: 701–10.

Malakar, Pushkar et al. 2016. "Insulin Receptor Alternative Splicing Is Regulated by Insulin Signaling and Modulates Beta Cell Survival." *Scientific Reports* 6(1): 31222. https://doi.org/10.1038/srep31222.

Malz, Mona et al. 2014. "Overexpression of Far Upstream Element (FUSE) Binding Protein (FBP)-Interacting Repressor (FIR) Supports Growth of Hepatocellular Carcinoma." *Hepatology* 60(4): 1241–50. https://doi.org/10.1002/hep.27218.

Maniatis, Tom, and Bosiljka Tasic. 2002. "Alternative Pre-MRNA Splicing and Proteome Expansion in Metazoans." *Nature* 418(6894): 236–43. https://doi.org/10.1038/418236a.

Martelly, William et al. 2021. "Synergistic Roles for Human U1 SnRNA Stem-Loops in Pre-MRNA Splicing." *RNA biology* 18(12): 2576–93.

Matlin, Arianne J., Francis Clark, and Christopher W.J. Smith. 2005. "Understanding Alternative Splicing: Towards a Cellular Code." *Nature Reviews Molecular Cell Biology* 6(5): 386–98.

Maude, Shannon L et al. 2014. “Chimeric Antigen Receptor T Cells for Sustained Remissions in Leukemia.” *The New England journal of medicine* 371(16): 1507–17.

———. 2016. “Efficacy and Safety of CTL019 in the First US Phase II Multicenter Trial in Pediatric Relapsed/Refractory Acute Lymphoblastic Leukemia: Results of an Interim Analysis.” *Blood* 128(22): 2801. https://www.sciencedirect.com/science/article/pii/S0006497119328022.

Meyer, Markus, and Josep Vilardell. 2009. “The Quest for a Message: Budding Yeast, a Model Organism to Study the Control of Pre-MRNA Splicing.” *Briefings in Functional Genomics and Proteomics* 8(1): 60–67.

Mimoso, Claudia A., and Karen Adelman. 2023. “U1 SnRNP Increases RNA Pol II Elongation Rate to Enable Synthesis of Long Genes.” *Molecular Cell* 83(8): 1264-1279.e10. https://doi.org/10.1016/j.molcel.2023.03.002.

Min, H, C W Turck, J M Nikolic, and D L Black. 1997. “A New Regulatory Protein, KSRP, Mediates Exon Inclusion through an Intronic Splicing Enhancer.” *Genes & development* 11(8): 1023–36.

Miro, Julie et al. 2015. “FUBP1: A New Protagonist in Splicing Regulation of the DMD Gene.” *Nucleic Acids Research* 43(4): 2378–89.

Modrek, Barmak, and Christopher Lee. 2002. “A Genomic View of Alternative Splicing.” *Nature Genetics* 30(1): 13–19.

Morgan, Jeffrey T, Gerald R Fink, and David P Bartel. 2019. “Excised Linear Introns Regulate Growth in Yeast.” *Nature* 565(7741): 606–11. https://doi.org/10.1038/s41586-018-0828-1.

Narayanan, Sujata, and Paul J Shami. 2012. “Treatment of Acute Lymphoblastic Leukemia in Adults.” *Critical reviews in oncology/hematology* 81(1): 94–102.

Ni, Xiaomin, Stefan Knapp, and Apirat Chaikuad. 2020. “Comparative Structural Analyses and Nucleotide-Binding Characterization of the Four KH Domains of FUBP1.” *Scientific Reports* 10(1): 13459. https://doi.org/10.1038/s41598-020-69832-z.

Nilsen, Timothy W., and Brenton R. Graveley. 2010. “Expansion of the Eukaryotic Proteome by Alternative Splicing.” *Nature* 463(7280): 457–63.

Orlando, Elena J. et al. 2018. “Genetic Mechanisms of Target Antigen Loss in CAR19 Therapy of Acute Lymphoblastic Leukemia.” *Nature Medicine* 24(10): 1504–6. http://dx.doi.org/10.1038/s41591-018-0146-z.

Pai, Athma A. et al. 2017. “The Kinetics of Pre-MRNA Splicing in the Drosophila Genome and the Influence of Gene Architecture.” *eLife* 6: 1–26.

Pai, Athma A et al. 2018. “Numerous Recursive Sites Contribute to Accuracy of Splicing in Long Introns in Flies.” *PLOS Genetics* 14(8): e1007588. https://doi.org/10.1371/journal.pgen.1007588.

Pan, Qun et al. 2008. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” *Nature Genetics* 40(12): 1413–15.

Papaemmanuil, E et al. 2011. “Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts.” *The New England journal of medicine* 365(15): 1384–95.

Piovesan, Allison et al. 2015. “Identification of Minimal Eukaryotic Introns through GeneBase, a User-Friendly Tool for Parsing the NCBI Gene Databank.” *DNA Research* 22(6): 495–503. https://doi.org/10.1093/dnares/dsv028.

———. 2019. “Human Protein-Coding Genes and Gene Feature Statistics in 2019.” *BMC*

Research Notes 12(1): 315. https://doi.org/10.1186/s13104-019-4343-8.

Plaschka, Clemens, Pei-Chun Lin, Clément Charenton, and Kiyoshi Nagai. 2018. "Prespliceosome Structure Provides Insights into Spliceosome Assembly and Regulation." *Nature* 559(7714): 419–22.

Quinn, Leonie M. 2017. "FUBP/KH Domain Proteins in Transcription: Back to the Future." *Transcription* 8(3): 185–92. https://doi.org/10.1080/21541264.2017.1293595.

Rabenhorst, Uta et al. 2009. "Overexpression of the Far Upstream Element Binding Protein 1 in Hepatocellular Carcinoma Is Required for Tumor Growth." *Hepatology* 50(4): 1121–29. https://doi.org/10.1002/hep.23098.

———. 2015. "Single-Stranded DNA-Binding Transcriptional Regulator FUBP1 Is Essential for Fetal and Adult Hematopoietic Stem Cell Self-Renewal." *Cell Reports* 11(12): 1847–55. https://www.sciencedirect.com/science/article/pii/S2211124715005744.

Rädle, Bernd et al. 2013. "Metabolic Labeling of Newly Transcribed RNA for High Resolution Gene Expression Profiling of RNA Synthesis, Processing and Decay in Cell Culture." *Journal of Visualized Experiments* (78): 1–11.

Rappsilber, Juri, Ursula Ryder, Angus I. Lamond, and Matthias Mann. 2002. "Large-Scale Proteomic Analysis of the Human Spliceosome." *Genome Research* 12(8): 1231–45.

Reymond Sutandy, F. X. et al. 2018. "In Vitro ICLIP-Based Modeling Uncovers How the Splicing Factor U2AF2 Relies on Regulation by Cofactors." *Genome Research* 28(5): 699–713.

Ronaghi, M et al. 1996. "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release." *Analytical biochemistry* 242(1): 84–89.

Roy, Meenakshi, Namshin Kim, Yi Xing, and Christopher Lee. 2008. "The Effect of Intron Length on Exon Creation Ratios during the Evolution of Mammalian Genomes." *Rna* 14(11): 2261–73.

Sanger, F, S Nicklen, and A R Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74(12): 5463–67.

Schmucker, Dietmar et al. 2000. "Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity." *Cell* 101(6): 671–84. https://www.sciencedirect.com/science/article/pii/S0092867400808788.

Schwalb, Björn et al. 2016. "TT-Seq Maps the Human Transient Transcriptome." *Science* 352(6290): 1225–28.

Seiler, Michael et al. 2018. "Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types." *Cell Reports* 23(1): 282-296.e4.

Sharma, Shalini et al. 2014. "Stem-Loop 4 of U1 SnRNA Is Essential for Splicing and Interacts with the U2 SnRNP-Specific SF3A1 Protein during Spliceosome Assembly." *Genes & development* 28(22): 2518–31.

Singh, Jarnail, and Richard A Padgett. 2009. "Rates of in Situ Transcription and Splicing in Large Human Genes." *Nature Structural & Molecular Biology* 16(11): 1128–33. https://doi.org/10.1038/nsmb.1666.

Singh, Ravindra N, Joonbae Seo, and Natalia N Singh. 2020. "RNA in Spinal Muscular Atrophy: Therapeutic Implications of Targeting." *Expert Opinion on Therapeutic Targets* 24(8): 731–43. https://doi.org/10.1080/14728222.2020.1783241.

Singh, Ravindra N, and Natalia N Singh. 2018. "Mechanism of Splicing Regulation of Spinal

Muscular Atrophy Genes BT - RNA Metabolism in Neurodegenerative Diseases." In eds. Rita Sattler and Christopher J Donnelly. Cham: Springer International Publishing, 31–61. https://doi.org/10.1007/978-3-319-89689-2_2.

Smith, Tom, Andreas Heger, and Ian Sudbery. 2017. "UMI-Tools: Modeling Sequencing Errors in Unique Molecular Identifiers to Improve Quantification Accuracy." *Genome research* 27(3): 491–99.

Sorek, Rotem, and Gil Ast. 2003. "Intronic Sequences Flanking Alternatively Spliced Exons Are Conserved between Human and Mouse." *Genome Research* 13(7): 1631–37.

Sotillo, Elena et al. 2015. "Convergence of Acquired Mutations and Alternative Splicing of CD19 Enables Resistance to CART-19 Immunotherapy HHS Public Access." *Cancer Discov* 5(12): 1282–95. http://cancerdiscovery.aacrjournals.org/ (November 5, 2019).

Sterner, Robert C, and Rosalie M Sterner. 2021. "CAR-T Cell Therapy: Current Limitations and Potential Strategies." *Blood Cancer Journal* 11(4): 69. https://doi.org/10.1038/s41408-021-00459-7.

SULLY, Gareth et al. 2004. "Structural and Functional Dissection of a Conserved Destabilizing Element of Cyclo-Oxygenase-2 MRNA: Evidence against the Involvement of AUF-1 [AU-Rich Element/Poly(U)-Binding/Degradation Factor-1], AUF-2, Tristetraprolin, HuR (Hu Antigen R) or FBP1 (Far-." *Biochemical Journal* 377(3): 629–39. https://doi.org/10.1042/bj20031484.

Tammer, Luna et al. 2022. "Gene Architecture Directs Splicing Outcome in Separate Nuclear Spatial Regions." *Molecular cell* 82(5): 1021-1034.e8.

Tellier, Michael, Isabella Maudlin, and Shona Murphy. 2020. "Transcription and Splicing: A Two-Way Street." *Wiley Interdisciplinary Reviews: RNA* 11(5): 1–25.

Terwilliger, T, and M Abdul-Hay. 2017. "Acute Lymphoblastic Leukemia: A Comprehensive Review and 2017 Update." *Blood cancer journal* 7(6): e577.

The Nobel Assembly at Karolinska Institute. 1993. "The Nobel Prize in Physiology or Medicine 1993." https://www.nobelprize.org/prizes/medicine/1993/press-release/.

Thiede, B, C Dimmler, F Siejak, and T Rudel. 2001. "Predominant Identification of RNA-Binding Proteins in Fas-Induced Apoptosis by Proteome Analysis." *The Journal of biological chemistry* 276(28): 26044–50.

Turunen, Janne J., Elina H. Niemelä, Bhupendra Verma, and Mikko J. Frilander. 2013. "The Significant Other: Splicing by the Minor Spliceosome." *Wiley Interdisciplinary Reviews: RNA* 4(1): 61–76.

Ule, Jernej, and Benjamin J Blencowe. 2019. "Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution." *Molecular cell* 76(2): 329–45.

Wachutka, Leonhard, Livia Caizzi, Julien Gagneur, and Patrick Cramer. 2019. "Global Donor and Acceptor Splicing Site Kinetics in Human Cells." *eLife* 8: 1–52.

Wahl, Markus C., and Reinhard Lührmann. 2015. "Snapshot: Spliceosome Dynamics I." *Cell* 161(6): 1474.e1-1474.e1. http://dx.doi.org/10.1016/j.cell.2015.05.050.

Wahl, Markus C., Cindy L. Will, and Reinhard Lührmann. 2009. "The Spliceosome: Design Principles of a Dynamic RNP Machine." *Cell* 136(4): 701–18.

Wang, Eric T. et al. 2008. "Alternative Isoform Regulation in Human Tissue Transcriptomes." *Nature* 456(7221): 470–76.

Wang, Jingxin, Peter G. Schultz, and Kristen A. Johnson. 2018. "Mechanistic Studies of a Small-

Molecule Modulator of SMN2 Splicing." *Proceedings of the National Academy of Sciences of the United States of America* 115(20): E4604–12.

Wilkinson, Max E, Clément Charenton, and Kiyoshi Nagai. 2020. "RNA Splicing by the Spliceosome." *Annual Review of Biochemistry* 89(1): 359–88. https://doi.org/10.1146/annurev-biochem-091719-064225.

Wudhikarn, Kitsada et al. 2021. "Interventions and Outcomes of Adult Patients with B-ALL Progressing after CD19 Chimeric Antigen Receptor T-Cell Therapy." *Blood* 138(7): 531–43.

Yoshida, Kenichi et al. 2011. "Frequent Pathway Mutations of Splicing Machinery in Myelodysplasia." *Nature* 478(7367): 64–69. https://doi.org/10.1038/nature10496.

Zhang, J, and Q M Chen. 2013. "Far Upstream Element Binding Protein 1: A Commander of Transcription, Translation and Beyond." *Oncogene* 32(24): 2907–16.

Zhang, Suyang et al. 2021. "Structure of a Transcribing RNA Polymerase II–U1 SnRNP Complex." *Science* 371(6526): 305–9. https://doi.org/10.1126/science.abf1870.

Zhang, Xiao-Ou et al. 2018. "The Temporal Landscape of Recursive Splicing during Pol II Transcription Elongation in Human Cells." *PLOS Genetics* 14(8): e1007579. https://doi.org/10.1371/journal.pgen.1007579.

Zhang, Yuanjiao, Jinjun Qian, Chunyan Gu, and Ye Yang. 2021. "Alternative Splicing and Cancer: A Systematic Review." *Signal Transduction and Targeted Therapy* 6(1): 78. https://doi.org/10.1038/s41392-021-00486-7.

Zheng, Wang et al. 2016. "Far Upstream Element-Binding Protein 1 Binds the 3′ Untranslated Region of PKD2 and Suppresses Its Translation." *Journal of the American Society of Nephrology* 27(9). https://journals.lww.com/jasn/fulltext/2016/09000/far_upstream_element_binding_protein_1_binds_the.14.aspx.

Zheng, Ying et al. 2020. "FUBP1 and FUBP2 Enforce Distinct Epigenetic Setpoints for MYC Expression in Primary Single Murine Cells." *Communications Biology* 3(1): 545. https://doi.org/10.1038/s42003-020-01264-x.

Zheng, Yuhuan, and W Keith Miskimins. 2011. "Far Upstream Element Binding Protein 1 Activates Translation of P27Kip1 MRNA through Its Internal Ribosomal Entry Site." *The International Journal of Biochemistry & Cell Biology* 43(11): 1641–48. https://www.sciencedirect.com/science/article/pii/S1357272511002068.

Zhou, Weixin et al. 2016. "Far Upstream Element Binding Protein Plays a Crucial Role in Embryonic Development, Hematopoiesis, and Stabilizing Myc Expression Levels." *The American journal of pathology* 186(3): 701–15.

9. Appendix

9.1. List of Figures

9.2. List of Tables

9.3. Abbreviations

3'ss	3' splice site
3'UTR	3' untranslated region
4sU	4-Thiouridine
5'ss	5' splice site
5'UTR	5' untranslated region
A3'SS	Alternative 3' splice site
A5'SS	Alternative 5' splice site
ARE	AU-rich elements
B-ALL	B cell acute lymphoblastic leukemia
BP	Branch point
BRET	Bio resonance energy transfer
CART	Chimeric antigen receptor T cell
ChIP-seq	Chromatin immunoprecipitation sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
cryo-EM	Cryogenic electron microscopy
CTD	C-terminal domain
Ctrl	Control
DNA	Deoxyribonucleic acid
DRB	5,6-Dichlorobenzimidazole 1-b-D-ribofuranoside
DTT	Dithiothreitol
e.g.	*exempli gratia*
ES	Exon skipping
ESE	Exonic splicing enhancers
FIR	FUBP1 interacting repressor
FUBP	FUSE-binding protein
FUSE	Far upstream element
GFP	Green fluorescent protein
GRO-seq	Global run-on sequencing
h	Hour
H3K36me3	Histone H3 lysine 36 tri-methylation
hnRNP	Heterogeneous ribonucleoprotein particle
iCLIP	Individual nucleotide crosslink and immunoprecipitation
IR	Intron retention
IRES	Internal ribosome entry site
ISS	Intronic splicing silencers
KD	Knockdown
kDa	Kilo Dalton
KH	K-homology
KHSRP	KH
KO	Knockout
LECA	Last eukaryotic common ancestor
LGG	Low grade glioma
LINE	Long interspersed nuclear elements
MPDZ	Multiple PDZ domain protein
mRNA	Messenger RNA
N-box^{mut}	N-box disrupting mutation
NGS	Next-generation sequencing
NLS	Nuclear localization signal
NMR	Nuclear magnetic resonance

nt	Nucleotides
PCR	Polymerase chain reaction
PolII	RNA polymerase II
pre-mRNA	Precursor messenger RNA
PSI	Percentage spliced in
PTC	Premature termination codon
py-tract	Polypyrimidine tract
RBP	Ribonucleoprotein
RIN	RNA integrity number
RNA	Ribonucleic acid
RPE1	Retina pigment epithelial 1
rRNA	Ribosomal RNA
RT-PCR	Reverse transcription and PCR
RT-qPCR	Real-time quantitative PCR
scFv	Single-chain variable fragment
seq	Sequencing
SF1	Splicing factor 1
SINE	Short interspersed nuclear elements
siRNA	Small interfering RNA
SMA	Spinal muscular atrophy
snRNA	Small nuclear RNA
snRNP	Small nuclear RBP
SR proteins	Serine arginine-rich proteins
tpm	Transcripts per million
U2AF2	U2 auxiliary factor
WT	Wild type

Danksagung

Diese Thesis enthält meine Arbeit, dennoch ist sie nicht mein alleiniger Verdienst, denn ohne die Kooperation und Unterstützung vieler Menschen wäre sie nicht möglich gewesen.

Zuerst möchte ich mich bei [redacted] für die Betreuung und die Möglichkeit, an diesem Projekt zu arbeiten bedanken. Seine Expertise, aber auch die offene, ehrliche Kommunikation, haben meine Zeit im Labor zu einer schönen und prägenden Erfahrung gemacht. Vielen Dank dafür, [redacted]!

Zu unserem FUBPl Team gehören allerdings auch noch weitere, einzigartige Wissenschaftler, allen voran [redacted]. Durch unseren Austausch habe ich viel bioinformatisches Wissen sammeln können. Auch in der Zeit, die wir unsere gemeinsame Publikation zusammengestellt haben, habe ich von Stefanie viel Halt erfahren. Mein Dank gilt auch [redacted], die durch ihre resiliente Art und ihre ausgezeichnete Wissenschaft das Projekt mitgeformt hat. Vielen Dank euch beiden!

Ich möchte mich auch bedanken bei allen Mitgliedern des [redacted] Labs, die mir und dem Projekt tagtäglich mit Rat und Tat zur Seite standen. [redacted] und [redacted] haben meine ersten Schritte als Doktorandin sowohl wissenschaftlich als auch menschlich begleitet. Auch allen Alumni und derzeitigen Mitgliedern des [redacted] Labs gilt mein Dank, denn nur mit einem rücksichtsvollen und offenen Team kann man diese Qualität an Wissenschaft erreichen, nach der wir tagtäglich streben. Danke für all die Hilfe, für jede konstruktive Kritik, aber auch für alle schönen Momente außerhalb des Labors, die wir miteinander teilen konnten.

Mein Dank gilt auch den Mitgliedern der [redacted] Gruppe und der [redacted] Gruppe, die durch enge Kollaboration und regen Austausch immer wieder neuen Input, eine neue Richtung und neuen Wind ins Projekt gebracht haben. Ebenfalls möchte ich mich bei meinen TAC Mitgliedern [redacted], [redacted], [redacted] und [redacted] für den Einsatz, die konstruktiven Diskussionen und all die Ratschläge bedanken. Ich möchte mich auch beim [redacted] und den [redacted] für eine erfolgreiche wissenschaftliche Zusammenarbeit bedanken. Danke auch an alle [redacted]-Mitglieder; gemeinsam haben wir selbst in Zeiten von COVID eine einzigartige Gemeinschaft aufgebaut.

Eine Thesis braucht aber nicht nur gute Wissenschaft, sondern benötigt auch einen starken Rückhalt. Ich möchte mich bei [redacted] bedanken, bei [redacted], und bei [redacted], die mich auf dem langen und oft auch beschwerlichen Weg durchs Studium unterstützt, aufgefangen und begleitet haben. Danke euch allen! Ohne euch wäre ich nicht der Mensch, der ich heute bin. Als letztes möchte ich mich bei [redacted] bedanken, meinem Partner und meinem Ruhepol. Deine Unterstützung und dein Verständnis sind unersetzlich. Danke, dass du diesen Weg mit mir teilst!