Deciphering the binding regulation of the core splicing factor U2AF65 using *in vitro* iCLIP

Dissertation
Zur Erlangung des Grades
Doktor der Naturwissenschaften

Am Fachbereich Biologie
Der Johannes - Gutenberg Universität Mainz

Reymond Sutandy
geb. am 13. Dezember 1988 in Klaten, Indonesien

Mainz, 10. September 2018

Dekan:

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung:

Sometimes our light goes out but is blown into flame by another human being. Each of us owes deepest thanks to those who have rekindled this light.

Albert Schweitzer

CONTRIBUTIONS

I would like to acknowledge the contributions of different people for this PhD project. I performed all *in vivo* and *in vitro* iCLIP experiments as well as validation experiments. Dr. Stefanie Ebersberger did all bioinformatics analysis of *in vivo* and *in vitro* iCLIP data as well as the machine learning analysis. Dr. Lu Huang conceived and implemented the binding model. Dr. Anke Busch was responsible for initial data processing of *in vivo* and *in vitro* iCLIP data. Maximilian Bach performed initial *in vitro* iCLIP co-factor assays with hnRNPC1. Dr. Hyun Seo Kang performed and analyzed ITC measurements. Dr. Jörg Fallmann calculated binding site accessibility score. Most data in this dissertation have been published as a research article in Sutandy et al., 2018.

TABLE OF CONTENTS

ZUSAMMENFASSUNG 1

SUMMARY 2

1. INTRODUCTION 3

1.1 Splicing and the spliceosome 3

1.2 Alternative splicing 5

1.3 Regulation of alternative splicing 6

1.3.1 Splice site recognition elements 6

1.3.2 Splicing enhancers and silencers 7

1.3.3 RNA secondary structure in splicing 9

1.4 U2AF65 and 3′ splice site definition 9

1.4.1 3′ splice site definition complex 9

1.4.2 U2AF65 structure and function 10

1.4.3 Regulation of U2AF65 binding in splicing 12

1.5 Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP) 13

1.5.1 Development of iCLIP 14

1.5.2 iCLIP workflow 16

1.6 *In vitro* techniques to study RNA-protein interactions 16

2. AIMS OF THE PROJECT 18

3. MATERIALS AND METHODS 19

3.1 Materials 19

3.1.1 Buffers 19

3.1.2 Reagents and disposables 21

3.1.3 Enzymes 22

3.1.4 Kits ... 23
3.1.5 Antibodies ... 24
3.1.6 Primers ... 25
3.1.7 siRNAs ... 29
3.2 Methods ... 31
3.2.1 Preparation of recombinant proteins ... 31
3.2.2 Preparation of in vitro transcript mix ... 32
3.2.3 In vitro iCLIP library preparation and sequencing ... 32
3.2.4 in vivo iCLIP library preparation and sequencing ... 36
3.2.5 Characterization of U2AF65 binding sites ... 36
3.2.6 Model-based estimation of in vitro K_d values ... 37
3.2.7 Model-based analysis of in vivo regulatory hotspots ... 38
3.2.8 K_d measurements by MST and ITC ... 39
3.2.9 Random Forests analysis ... 39
3.2.10 Analysis of in vitro iCLIP co-factor assays ... 40
3.2.11 Knockdown of RBPs ... 40
3.2.12 Western blot ... 41
3.2.13 Minigene reporter assays ... 41
3.2.14 In vivo splicing assays ... 42
4. RESULTS ... 43
4.1 Establishment of the *in vitro* iCLIP protocol ... 43
4.2 U2AF65^{RRM12} resembles full-length U2AF65 binding *in vitro* ... 48
4.3 The *in vitro* U2AF65^{RRM12} binding landscape differs from *in vivo* binding ... 49
4.4 Transcript-wide measurement of U2AF65 binding site affinities ... 50
4.5 U2AF65 binding is heavily regulated *in vivo* ... 56

4.6 Machine learning identifies RBPs as potential U2AF65 regulators 59
4.7 Validation of predicted U2AF65 regulatory events *in vitro* 61
4.8 *In vivo* regulation by hnRNPC can be recapitulated *in vitro* 63
4.9 *PTBP2* exon 10 alternative splicing regulation by PTBP1 and FUBP1 66
4.10 Relevance of predicted regulation to splicing decision *in vivo* 69
5. DISCUSSION 73
5.1 *In vitro* iCLIP measures binding site affinity in a natural RNA context 73
5.2 Calibration of the *in vivo* iCLIP signal 74
5.3 U2AF65 binding is stabilized at 3′ splice site and cleared in intronic region 74
5.4 Identification of U2AF65 regulators 75
5.5 *In vitro* iCLIP disentangles complex regulatory mechanisms *in vivo* 77
6. REFERENCES 79
7. APPENDIX 89
7.1 List of Figures 89
7.2 List of Tables 90
7.3 Abbreviations 91

ZUSAMMENFASSUNG

Durch Bildung unterschiedlicher reifer mRNAs aus einzelnen Genen erhöht alternatives Spleißen die proteomische Vielfalt in Eukaryonten. RNA-Protein Interaktionen organisieren die Regulation des alternativen Spleißens in einem mehrstufigen Prozess, der die Regulation der Bindung des essentiellen Spleiß-Faktors U2AF65 im Rahmen der Definition der 3‘ Spleißstelle umfasst. Zur Untersuchung der U2AF65-Bindung und deren Regulation durch andere RNA bindende Proteine (RBP) haben wir ‘*in vitro* iCLIP’ entwickelt, das ermöglicht RNA-Protein Interaktionen im definierten System zu erfassen. Mit Hilfe von rekombinanten Proteinen und einer Reihe von *in vitro* Transkripten wurde *in vitro* iCLIP zur Transkript-weiten Messung von U2AF65 Bindungsaffinitäten verwendet. Auf den Affinitätsdaten basierende Mathematische Modellierung ermöglichte den Vergleich der *in vitro* und *in vivo* U2AF65 Bindungs-Muster und verschaffte uns umfangreiche Informationen über die Transkript-weite *in vivo* Regulation der Bindung. Wir haben herausgefunden, dass U2AF65 Bindung umfassend reguliert ist, einschließlich der Stabilisierung von U2AF65 an 3‘ Spleißstellen und Entfernung des Proteins von intronischen Regionen. Darüber hinaus wurde auf RBP Sequenzmotiven basiertes, maschinelles Lernen verwendet um RBP zu identifizieren, die die U2AF65 Bindung an regulierten Bindestellen potentiell modulieren. Dadurch haben wir eine Handvoll bekannter und neuer Interaktoren der U2AF65 Bindung identifiziert. Zusätzliche *in vitro* Validierung offenbarte, dass hnRNPC, PTBP1 und PCBP1 hauptsächlich als Binde-Suppressoren in Erscheinung treten, während FUBP1 und CELF6 die Bindung von U2AF65 verstärken. Knock-down Experimente dieser RBP deuteten darauf hin, dass die *in vitro* Modulierung der U2AF65 Bindung durch diese RBP den Ausgang des alternativen Spleißens *in vivo* beeinflusst. Zusammenfassend betrachtet liefert *in vitro* iCLIP eine Plattform, welche die Charakterisierung von RNA-Protein Interaktionen im vereinfachten System ermöglicht und zur Unterstützung der Interpretation komplexer *in vivo* Binde-Daten genutzt werden kann.

SUMMARY

Alternative splicing increases proteome diversity in eukaryotes by generating different variants of mature mRNA from single genes. RNA-protein interactions orchestrate the regulation of alternative splicing in a multi-step manner, including binding modulation of the core-splicing factor U2AF65 in 3' splice site definition. To study U2AF65 binding and its modulation by other RNA binding proteins (RBPs), we developed '*in vitro* iCLIP' that enables capturing RNA-protein interactions in a defined system. By using recombinant proteins and a set of *in vitro* transcripts, we applied *in vitro* iCLIP to measure U2AF65 binding site affinities in a transcript-wide manner. Mathematical modeling based on the affinity data allowed comparison of the *in vitro* and *in vivo* U2AF65 binding landscapes, and provided us with comprehensive information on the transcript-wide *in vivo* binding regulation. We found that U2AF65 binding is extensively regulated, including stabilization at 3' splice sites and clearance of binding in intronic regions. Furthermore, a machine learning approach based on RBP sequence motifs was used to identify RBPs that potentially modulate U2AF65 binding at the differentially regulated sites. As a result, we identified a handful of known and novel regulators of U2AF65 binding. Further *in vitro* validation revealed that hnRNPC, PTBP1, and PCBP1 mainly act as suppressors, whereas FUBP1 and CELF6 enhance U2AF65 binding. Knockdown experiments of the RBPs indicated that *in vitro* U2AF65 binding modulations by these RBPs affect the alternative splicing outcome *in vivo*. In conclusion, *in vitro* iCLIP provides a platform that allows characterization of RNA-protein interactions in a simplified system and can be used to aid in the interpretation of complex *in vivo* binding data.

1. INTRODUCTION

1.1 Splicing and the spliceosome

To form mature mRNAs used as templates for protein translation, most eukaryotic genes go through several steps of post-(co-)transcriptional RNA processing. Among these steps, RNA splicing is responsible for early RNA processing to remove the intervening sequences (introns) from pre-mRNA and ligate the expressed sequences (exons) together to form mRNA. During splicing, accuracy in defining the exons from the introns (exon-intron boundaries) is crucial to produce functionally correct transcripts that drive specific cellular programs. This role is accomplished by one of the largest ribonucleoprotein complexes in the cell, known as the spliceosome (Will & Lührmann, 2011; Matera & Wang, 2014).

There are two types of spliceosome in eukaryotes based on the type of their target introns: a rare class named the U12-dependent spliceosome, and the more abundant class called the U2-dependent spliceosome. This project focuses only on the latter class, the U2-dependent spliceosome, which is responsible for removing U2-type introns that are present in most eukaryotic pre-mRNAs. This type of spliceosome consists of five primary uridine-rich subunits (U1, U2, U5, U4/U6) called small nuclear ribonucleoprotein (snRNP) complexes and numerous other splicing factors that orchestrate accurate yet flexible control of splicing (Shi, 2017; Matera & Wang, 2014). Each snRNP comprises snRNAs that are synthesized and assembled together with proteins to form a ribonucleoprotein complex through an extensive process that occurs in several different cellular compartments (Matera & Wang, 2014). The final assembly of the snRNPs to form the spliceosome happens in the nucleus during the initial step of splicing.

Splicing occurs in a step-wise manner, starting with the recognition of exon-intron boundaries, known as 5′ and 3′ splice sites. U1 snRNP is responsible for recognizing the 5′ splice site, whereas U2 and its auxiliary factors define the 3′ splice site (Shi, 2017; Will & Lührmann, 2011). Recognition of both splice sites marks the beginning and initiates further catalytically active processes of splicing that generally involve two consecutive transesterification reactions (**Figure 1**). Evolutionary, these processes resemble the removal of mobile genetic elements called group II introns that exist in bacteria, mitochondria, and

chloroplasts but are absent in the eukaryotic nuclear genome (Papasaikas & Valcárcel, 2016), suggesting that they might be derived from the same origin.

Differences in the eukaryotic genome architecture introduce variation in splicing. The most common form of this variation happens in the formation of pre-spliceosomal complexes. In yeast, the pre-spliceosomes are formed spanning the introns via the interaction between U1 and U2 snRNPs (Matera & Wang, 2014). However, in higher eukaryotes where exons are often separated by large introns (up to 1000 kb), it is an arduous task for the splicing machineries to form the pre-spliceosomes spanning the introns. Instead, the interaction between U1 and U2 snRNPs initially happens across the exons, which are typically much shorter in length than the introns, through a process known as exon definition (Matera & Wang, 2014; Berget, 1995). In a poorly understood process, the U1 and U2 snRNPs undergo subsequent rearrangement to eventually form complex spanning the introns.

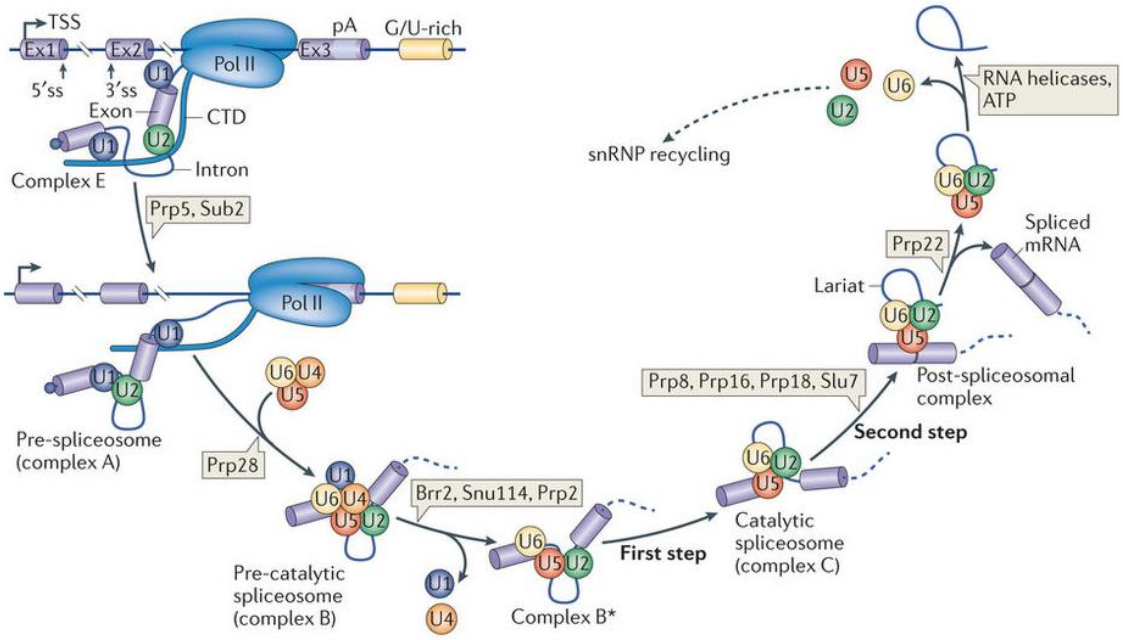


Figure 1. Stages of splicing in eukaryotes. Splicing starts with the formation of complex E through recognition of 5′ splice site by U1 (facilitated by the CTD of Pol II), and 3′ splice site by U2 and its auxiliary factors in an ATP-independent manner. The further consecutive processes will then require ATP to initiate the formation of pre-spliceosome (complex A) through an interaction between U1 and U2 snRNPs. The pre-assembled U5-U4/U6 complex is then recruited to form the pre-catalytic spliceosome (complex B). The complex B will go through several conformational changes to form a catalytically active complex B* that leads to the release of U1 and U4 snRNPs, and the formation of U2-U6 snRNA structure. Two catalytic steps of splicing are then started by first releasing exon 1 (Ex1) and forming the intron-exon 2 (Ex2) lariat intermediate (complex C). In the second catalytic step, exon 2 is released to form the intron lariat. The released exons are then ligated together to form the spliced mRNA, and the U5, U2 and U6 snRNPs are released to be recycled for the next splicing event. This figure was modified from Matera & Wang, 2014.

1.2 Alternative splicing

In higher eukaryotes, some of their genes go through alternative splicing, where a single pre-mRNA can be processed into different variants of mature mRNA. This process provides a possibility to expand their proteome diversity with a relatively limited number of genes in their genome. The extent of alternatively spliced genes varies between species and seems to be correlated with their complexity (Lee & Rio, 2015). In humans, more than 95% of genes are alternatively spliced. This proportion is higher than other vertebrates with a relatively similar number of genes; for example, in the mouse and the worm, only ~63% and ~25% of genes are alternatively spliced, respectively (Lee & Rio, 2015).

Alternative splicing is crucial for the maintenance of cell identity in multicellular organisms. Changes in the alternative splicing of a certain group of genes have been shown to be unique signatures for specific cellular functions, such as differentiation and cell proliferation (Fiszbein & Kornblihtt, 2017; Dominguez et al., 2016). These signatures are produced by different forms of mature mRNA, known as isoforms, at a specific time and cell type. Isoforms produced from the same gene can have a specific, substituting, or even opposite function, which drives specific cellular programs. Therefore, failure in executing the correct alternative splicing program leads to the misregulation of cellular functions and has been linked to the development of diseases including cancer (Oltean & Bates, 2014; Scotti & Swanson, 2016).

1.3 Regulation of alternative splicing

In the majority of cases, regulation of alternative splicing happens quite early during splice site definition. However, evidence shows that this regulation can also occur at different stages of splicing (Chen & Manley, 2009). In addition, because many transcripts are spliced co-transcriptionally, the contribution of transcription-coupled regulations, such as transcription rate (Dujardin et al., 2013) and chromatin marks (Luco et al., 2010), has recently emerged as interesting aspects to consider in the regulation of alternative splicing. This multitude of factors that contribute to the regulation makes alternative splicing an extremely dynamic process in the cell.

1.3.1 Splice site recognition elements

An accurate splicing program relies on the recognition of specific sequences in the RNA, known as splice site recognition elements. There are three main splice site recognition elements that are important for splice site definition: 5′ splice sites, 3′ splice sites and branch point sequences (**Figure 2**). In U2-dependent splicing, the 5′ splice sites are marked by a GU dinucleotide and recognized by U1 snRNPs. The 3′ splice sites consist of a branch point sequence and polypyrimidine (Py) tract, which is followed by a AG dinucleotide. The 3′ splice site regions initially are recognized by U2 auxiliary factors (U2AF) via their binding to the Py tract and AG dinucleotide, which then recruit U2 snRNPs to the branch point sequences to initiate splicing. In higher eukaryotes, splice site recognition elements are more degenerative than those in lower eukaryotes such as yeast (Lee & Rio, 2015). This characteristic allows for a more dynamic splice site definition of the individual exon, and is the primary reason for the higher proportion of alternatively spliced genes in higher eukaryotes than that in yeast where most genes are constitutively spliced.

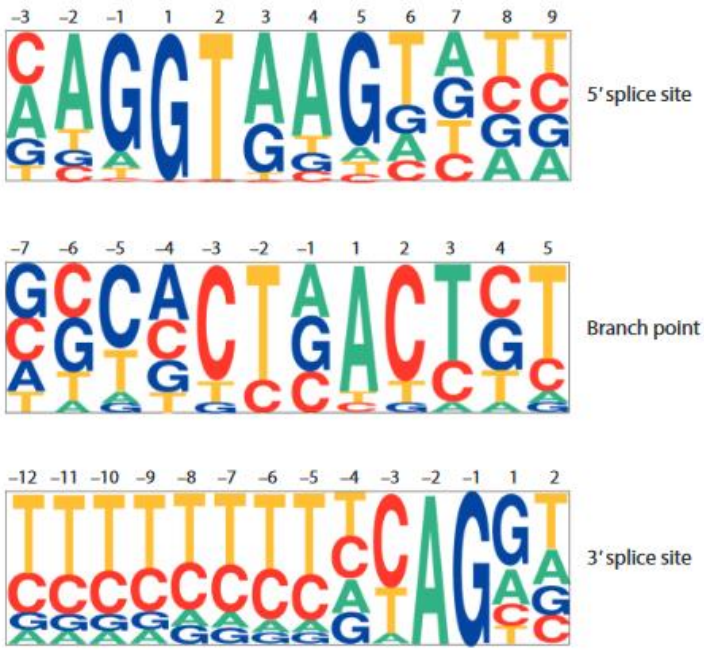


Figure 2. Pictogram of the U2-dependent splice site recognition elements. For the 5′ splice site, label 1 is the first nucleotide of the intron. For the branchpoint, label 1 is the branchpoint residue. For the 3′ splice site label -1 is the last nucleotide of the intron. The size of the letter in the pictogram represents the frequency of the corresponding base in each position. This figure was modified from Lee & Rio, 2015.

1.3.2 Splicing enhancers and silencers

The interactions between RNA and RNA binding proteins (RBPs) orchestrate the major part of alternative splicing regulatory events. Several RBPs that have functional effects on splicing regulation are often classified as splicing factors. These RBPs regulate splicing by binding to specific sequences on the RNA called *cis*-elements. Depending on the location and their functional effect in splicing, *cis*-regulatory elements can be classified into four categories: exonic splicing enhancers (ESE), exonic splicing silencers (ESS), intronic splicing enhancers (ISE), and intronic splicing silencers (ISS) (**Figure 3**, Chen & Manley, 2009).

The binding of specific RBP to *cis*-elements affects the outcome of alternative splicing in a multitude of ways. The binding of RBPs to splicing enhancers improves splicing by stabilizing recognition of the splice sites. Serine/arginine-rich (SR) proteins are examples of

splicing factors that bind to splicing enhancers and then recruit U1 and U2AF to the 5′ and 3′ splice sites, respectively, via direct protein-protein interaction (Bourgeois et al., 1999; Zuo & Maniatis, 1996). Conversely, splicing silencers recruit RBPs and their binding interaction inhibits recognition of the splice site. This inhibition most commonly occur in two ways: (1) direct competition between the RBPs and spliceosome components, where the locations of the splicing silencers overlap or in close proximity to the splice site recognition elements; (2) stearic hindrance to the activators binding at the splicing enhancers, preventing the recruitment of spliceosome components by the activators (Chen & Manley, 2009). An example of RBPs that commonly bind to splicing regulatory silencers are proteins from the heterogenous nuclear RNP (hnRNP) family (Singh et al., 1995; Zarnack et al., 2013).

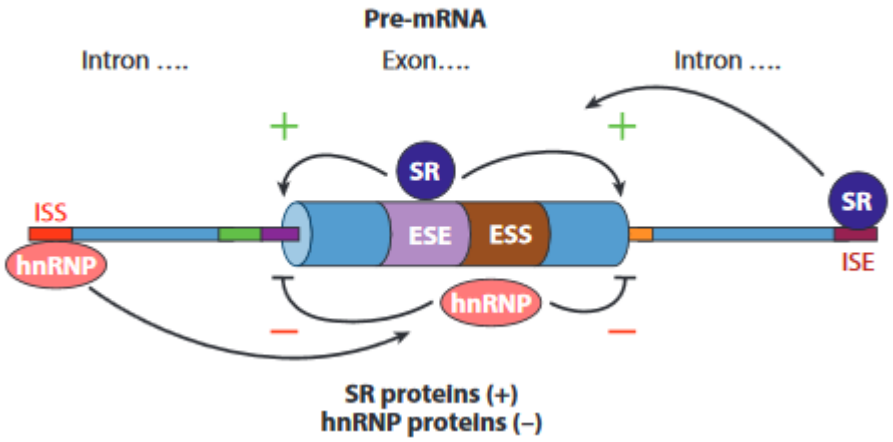


Figure 3. Schematic of the splicing regulatory events. Splicing factors such as SR proteins can bind to the splicing regulatory enhancers both in the intron (ISE) and exon (ESE) to enhance the splice site recognition. Inversely, another factors such as hnRNP proteins can inhibit the splice site recognition by binding to the splicing regulatory silencers in the exon (ESS) and intron (ISS). Combinatorial effect of these regulatory events will decide the final outcome of the alternative splicing. This figure was modified from Lee & Rio, 2015.

Although many splicing factors have a general effect on the direction of alternative splicing regulation, in some cases this regulatory effect is position dependent. A single splicing regulator can promote either splicing inhibition or activation depending on the binding site location relative to the regulated exon. Such cases have been shown for many splicing factors such as NOVA1 and hnRNPH (Dredge & Darnell, 2003; Dredge et al., 2005;

Schaub et al., 2007; Caputi & Zahler, 2001). In addition, the final splicing decision represents the combinatorial effect of several splicing regulatory events that happen around the alternative exons. Most of these functional regulatory events occur within 200–300 nucleotides away from the splice sites. Nevertheless, a recent study has shown that remote regulatory events (>500 nucleotide) are as important to consider when evaluating the final splicing decision (Lovci et al., 2013; Fu & Ares, 2014).

1.3.3 RNA secondary structure in splicing

Most steps of splicing are based on the recognition of specific sequences in the pre-mRNA via base pairing with snRNAs or the binding of RBPs. Therefore, internal base pairing in the pre-mRNA that forms certain secondary structures can inhibit the recognition of *cis-elements*. Several studies have shown the inhibitory effect of RNA secondary structures. For example, the inhibition of the U1-mediated 5′ splice sites recognition; and both the U2 snRNAs- and U2AF-mediated 3′ splice sites recognitions by the RNA secondary structures (Warf et al., 2009; Sirand-Pugnet et al., 1995; Singh et al., 2007). Inversely, certain folding of RNA structures can also enhance the splicing process. This enhancement usually happens when RNA folding is formed between splicing regulatory elements, thereby bringing them into a closer proximity (Warf & Berglund, 2010). A handful of RBPs have been described as regulating alternative splicing via modulation of RNA secondary structures, such as MBNL1, PTBP1, and hnRNPA1 (Pascual et al., 2006; Oberstrass et al., 2005; Blanchette & Chabot, 1999).

1.4 U2AF65 and 3′ splice site definition

1.4.1 3′ splice site definition complex

The initial step of 3′ splice site recognition is executed by a ternary complex of U2AF and a splicing factor, SF1. U2AF is a heterodimer that consists of two subunits, U2AF65 and U2AF35. U2AF65 recognizes the Py tract preceding 3′ splice site, whereas U2AF35 binds to a AG dinucleotide at 3′ splice site (Chatrikhi et al., 2016). SF1 recognizes the third 3′ splice site recognition elements, the branch point sequence. In addition to their binding to the RNA,

U2AF65, U2AF35, and SF1 form a complex via direct protein-protein interaction through a specific domain known as the U2AF homology motif (UHM). The formation of the U2AF65-U2AF35-SF1 complex marks the early E complex of splicing. Upon the formation of this complex, U2 snRNP is recruited in an ATP-dependent manner. U2 snRNP then displaces SF1 binding to the branch point, and its subunit SF3B155 substitutes the SF1 interaction with the U2AF heterodimer to initiate further splicing (Chatrikhi et al., 2016). The formation of the U2AF65-U2AF35-SF1 complex is especially important in higher eukaryotes because the branch point sequence is very degenerate, and therefore, auxiliary factors are required to aid in the recruitment of U2 snRNP to the 3′ splice site. In humans, ~88% of the 3′ splice sites are recognized by U2AF (Shao et al., 2014).

1.4.2 U2AF65 structure and function

U2AF65 harbors three RNA recognition motif (RRM) domains. The first two N-terminal RRM domains (RRM1 and RRM2) are canonical RRM domains that are responsible for U2AF65 binding to the Py tract. The third RRM domain in the C-terminal region, which is the UHM domain of U2AF65, prefers to bind to tryptophan-containing linear peptide motifs called UHM ligand motifs in several nuclear proteins (Corsini et al., 2009; Voith von Voithenberg et al., 2016). Although the RRM3 domain of U2AF65 is relatively conserved, *in vitro* studies suggest that this domain is dispensable for the function of U2AF65 to initiate splicing. Therefore, many *in vitro* studies of U2AF65 use the minimal construct of U2AF65 that contains only the RRM1 and RRM2 domains (Mackereth et al., 2011; Banerjee et al., 2004). However, *in vivo*, the RRM3 domain has an important role in U2AF65 splicing function for mediating the interactions with SF1 and possibly other splicing factors (Selenko et al., 2003; Banerjee et al., 2004).

Based on its domains arrangement, U2AF65 is present in two different conformations (Mackereth et al., 2011). A “closed” conformation is formed when RRM1 interacts with RRM2 of U2AF65, where only RRM2 is available for binding to RNA. This conformation can be reversed to the “open” conformation when a high affinity binding site on the RNA interacts with RRM2 and thereby competes for the binding with RRM1. The high affinity

binding site will then open the interaction between RRM1 and RRM2, making both RRM domains available for binding to the RNA (**Figure 4**). This multi-domain conformational switch of U2AF65 is thought to be important for the ability of U2AF65 to differentiate the strengths of the Py tract.

Although U2AF65 is known to recognize a wide variety of Py tracts, the basis of how the sequence content of the Py tract defines U2AF65 binding affinity is rather puzzling. A systematic study on the effect of different Py tract lengths and content revealed that 11 consecutive uracils have the strongest ability to initiate splicing (Coolidge et al., 1997). A decrease in Py tract length or an interruption by purine bases reduces Py tract strength while still maintaining the ability to initiate splicing. However, in the case of weaker Py tracts, they must be located in close proximity to the AG dinucleotide at the 3′ splice site. Structural data of U2AF65 show that seven uracils directly interact with residues from RRM1 and RRM2, and additional two uracils in the middle of the Py tract interact with the linker region between RRM1 and RRM2 to establish binding to RNA (Sickmier et al., 2006; Agrawal et al., 2016). *In vitro*, it has been shown that a long uninterrupted Py tract with at least 9 pyrimidines is enough to form the stable open conformation of U2AF65 (Mackereth et al., 2011). Overall, the length of the Py tract is a major contributing factor to the strength of the 3′ splice site.

In the cell, U2AF65 functions as a heterodimer with U2AF35. However, the interdependency between these two subunits for their role in splicing is not clear. *In vitro* studies have shown that U2AF65 alone can complete the splicing process in the absence of U2AF35 (Zamore et al., 1992; Valcárcel et al., 1996). Nevertheless, several roles of U2AF35 in stabilizing U2AF65 binding have been demonstrated *in vivo* in the case of weak splice site recognition. This stabilization happens either via direct interaction or by mediating U2AF65 interaction with splicing enhancers such as SR proteins (Wu & Maniatis, 1993; Wu et al., 1999). A recent structural study of the U2AF heterodimer showed that the enhancement of U2AF65 binding in weak splice sites may also occur via the promotion of open conformation U2AF65 in the presence of U2AF35 (Voith von Voithenberg et al., 2016).

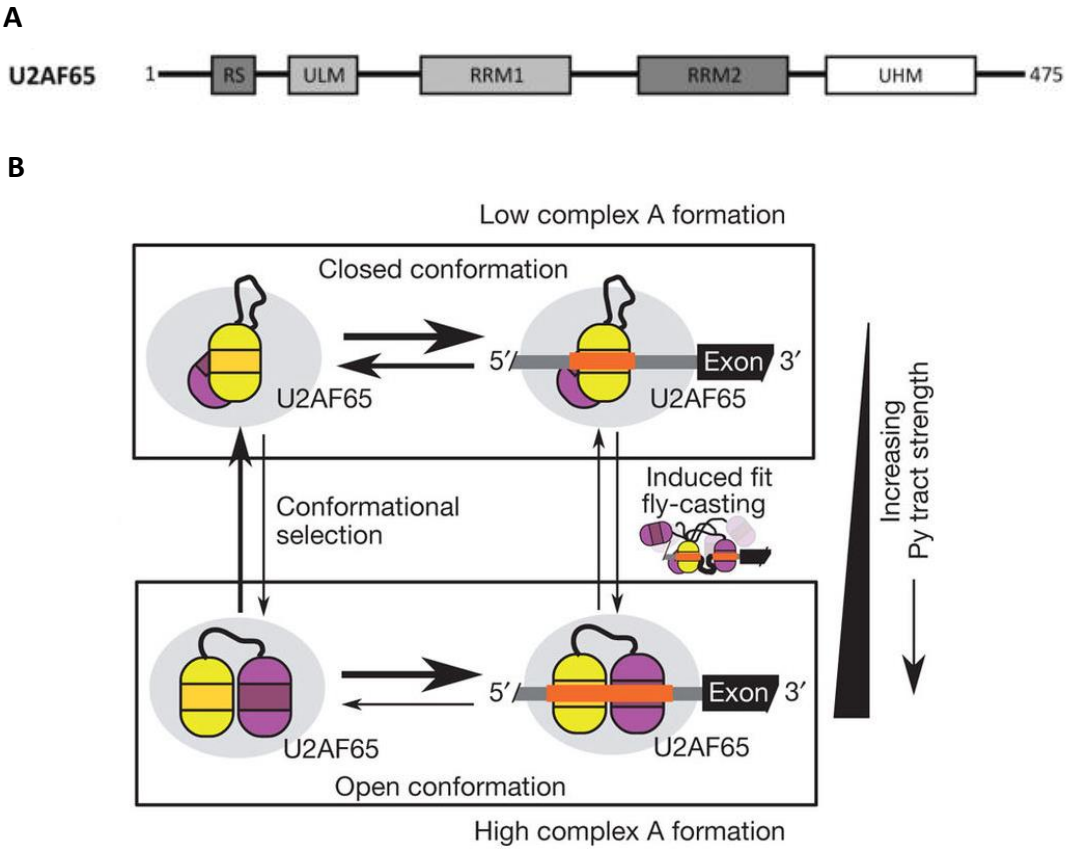


Figure 4. Domain architecture and conformation of U2AF65 (A) U2AF65 domain composition. This figure was modified from Voith von Voithenberg et al., 2016. **(B)** Equilibrium of U2AF65 multidomain conformations. In equilibrium, U2AF65 present in both "open" and "closed" conformations with the tendency toward the "closed conformation". Upon binding to RNA, the equilibrium changes according to the Py tract strength of the bound RNA. The stronger the Py tract affinity, the closer the equilibrium moves toward "open conformation", and vice versa. This figure was modified from Mackereth et al., 2011.

1.4.3 Regulation of U2AF65 binding in splicing

U2AF65 binding is the prime target for alternative splicing regulation at the 3′ splice site. Several different splicing factors have been shown to modulate U2AF65 binding in a multitude of ways, thereby affecting the alternative splicing decision. Direct competition for Py tract binding is one of the most common regulations of U2AF65 binding. This is mostly derived from the promiscuous binding preference of U2AF65 toward different Py tract compositions, and thus, often overlaps with the binding motifs of several different RBPs. Among them, hnRNPC has been shown to generally compete with U2AF65 for a uridine-rich

region (Zarnack et al., 2013). Another splicing factor, PTBP1, binds to overlapping regions with U2AF65 and changes the splicing pattern of several genes (Saulière et al., 2006), suggesting that direct competition between these two factors may occur. A specific type of indirect competition was observed between MBNL1 protein and U2AF65 (Warf et al., 2009). These proteins compete for the 3′ splice site of cardiac troponin T (*cTNT*) intron 4 that can form mutually exclusive RNA structure. The stem loop structure of this region is bound by MBNL1, preventing the interaction with U2AF65, which prefers to bind to the single-stranded version of this region.

SR proteins have been known as splicing factors that bind to the exonic region and recruit U2AF65 to the splice sites with a weak Py tract (Graveley et al., 2001). The interaction between U2AF65 and the SR proteins is directly mediated by U2AF35. Mutating a weak Py tract to a stronger one has been shown to relieve the requirement for the splicing enhancer to complete splicing. A more complex regulation of U2AF65 binding was observed in the case of hnRNPA1 (Tavanez et al., 2012). This RBP can stabilize and form a ternary complex with U2AF65-U2AF35 in an AG-containing Py tract. Intriguingly, hnRNPA1 shows an opposite effect by displacing U2AF65 binding in a non-AG-containing Py tract. The detailed mechanism of this regulation is still unclear, but such regulation may play a role in the proofreading mechanism to ensure U2AF65 binding enrichment at 3′ splice site.

Another flavor in U2AF65 binding modulation occurs via post-translational modification of U2AF65. A demethylase and hydroxylase, Jumonji domain–containing 6 protein (JMJD6), has been shown to regulate alternative splicing via hydroxylation of several RS domain–containing splicing factors, including U2AF65 (Yi et al., 2016). A recent study showed that at least 30 alternative splicing events are regulated by the JMJD6-dependent U2AF65 hydroxylation, revealing yet another layer of U2AF65 regulations that has yet to be extensively investigated.

1.5 Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP)

The importance of RNA-protein interactions in the regulation of diverse cellular processes has spurred the development of techniques to study such interactions. Recently,

different variants of the high-throughput-based technique were developed to capture RNA-protein interactions in a genome-wide manner. Among them, iCLIP allows for investigation of RBP binding locations in detailed single nucleotide resolution, which provides better precision for studying RNA-protein interactions.

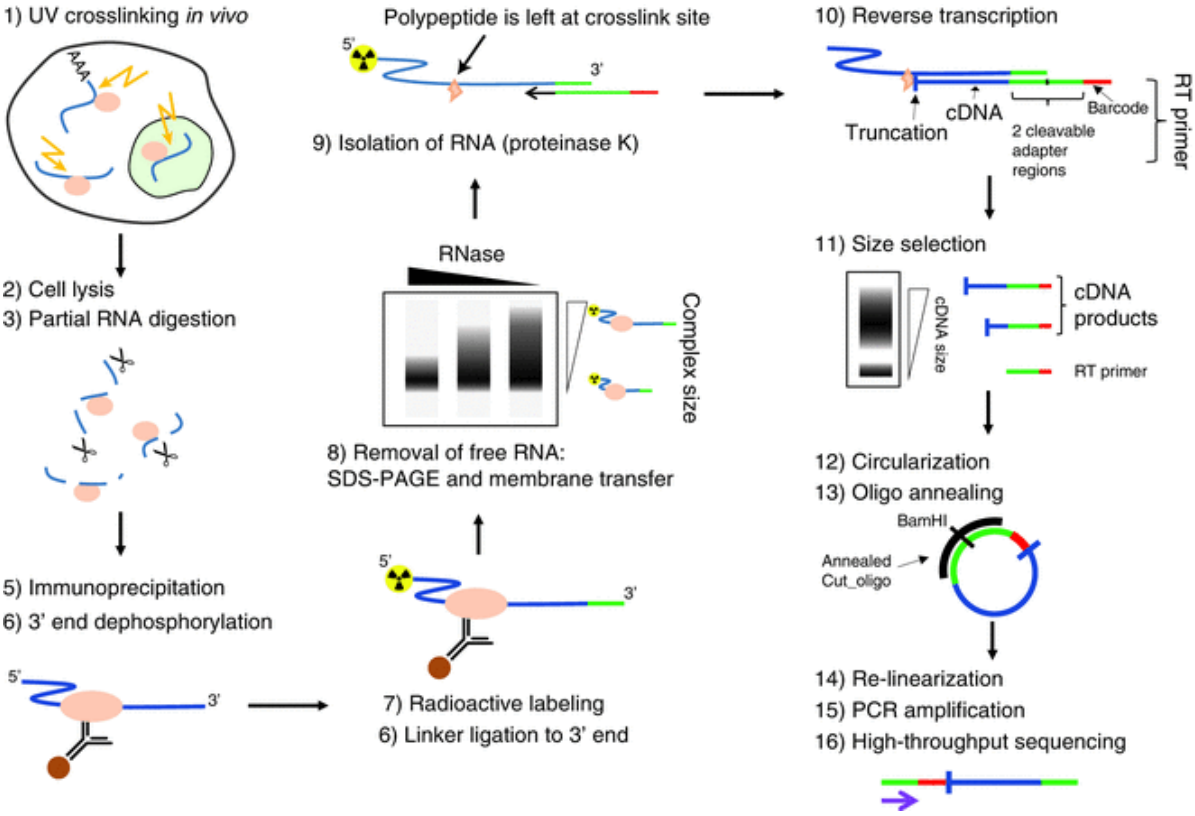
1.5.1 Development of iCLIP

One of the breakthroughs in studying RNA-protein interactions was the development of crosslinking and immunoprecipitation (CLIP) technology. With CLIP technology, living cells can be crosslinked by UV irradiation (254 nm), inducing the formation of covalent bonds between RNA and the interacting protein (Darnell, 2010; Ule et al., 2003). This crosslinking is followed with immunoprecipitation of a specific RBP of interest, allowing for co-precipitation of the interacting RNAs. Next, these RNAs can be released and extracted by digesting the RBP and then sequenced to reveal their identities and map the specific locations of RBP interactions. To improve the power of CLIP technology, a subsequent CLIP-based technique known as CLIP-Seq (also known as HITS-CLIP) was developed by employing high-throughput sequencing to identify the target RNAs (Licatalosi et al., 2008). Compared with the prior CLIP technique, CLIP-Seq has increased the output of the technique by more than 1000 folds, which enables a more robust generalization on the behavior of the RNA-protein interactions (Darnell, 2010).

Since the development of CLIP-Seq, research in this area has focused on enhancing the resolution for locating RBP binding sites. This, in particular, was necessary because the original CLIP-Seq mapping relies on grouping overlapping reads into clusters to identify binding site locations, which results in relatively broad binding site identification (König et al., 2010). In addition, the sensitivity of the CLIP-Seq protocol also suffers from the requirement of reverse transcriptase to bypass the remnant of small peptides that covalently attaches to RNA at the crosslinked site to identify the binding site location. In particular, the majority of cDNA produced during reverse transcription in the CLIP protocol has been shown to be truncated directly before the crosslinked site (Urlaub et al., 2002). Taking advantage of this truncation to locate the binding site, iCLIP was developed based on the CLIP-Seq

technology with improved sensitivity (König et al., 2010). Because this truncation often happens at the nucleotide adjacent to the crosslinked site, iCLIP used this positional information to pinpoint the binding site location of interacting RBPs down to single nucleotide resolution.

A



B

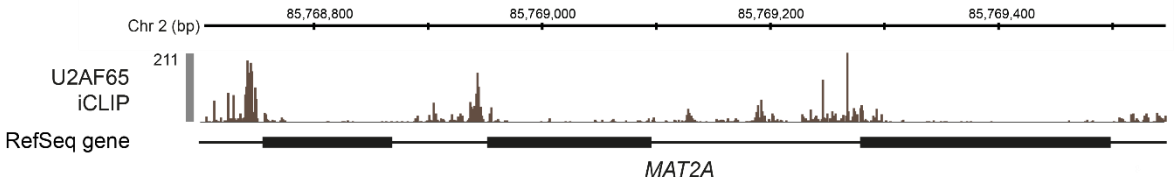


Figure 5. Overview of iCLIP protocol to study RNA-protein interactions. (A) Workflow of iCLIP library preparation. This figure was modified from Sutandy et al., 2016 **(B)** Data output of U2AF65 iCLIP. Peaks represent the reads mapped to the specific locations of the RNA-protein interaction. iCLIP shows the enrichment of U2AF65 binding at 3′ splice sites. Supporting previously described function of U2AF65 in 3′ splice site definition. Boxes represent exons, while lines represent introns.

1.5.2 iCLIP workflow

To crosslink the RNA-protein complexes, the iCLIP protocol starts with UV irradiation of living cells/tissues (**Figure 5A**, Sutandy et al., 2016). The samples are then lysed to release the crosslinked complexes, followed by a partial RNase digestion. This fragmentation of the interacting RNAs is important to obtain the range of RNA fragment size that is optimal for the downstream high-throughput sequencing application. The RNA-protein complexes are isolated by immunoprecipitation with antibody targeting the RBP of interest. Stringent washing procedures can be applied to reduce unspecific binding after the immunoprecipitation. A DNA linker is ligated to the RNAs to provide the specific sequence for the downstream steps of library preparation. The RNAs are then labeled with hot P^{32} and visualized with SDS-PAGE and membrane transfer, followed by autoradiography. The visualization allows for quality checking of the previous steps in the library preparation and the removel of noncrosslinked RNA. Interacting RNAs are extracted from the membrane by proteinase K digestion and further used for the input of cDNA synthesis with reverse transcription primers targeting the linker region. The resulting cDNAs are size selected, followed by circularization and relinearization to provide a platform for the PCR primers to anneal at both 3′ and 5′ ends of the cDNA. The library is further amplified and sequenced in the high-throughput sequencing platform. The sequencing output from the library can be mapped to a specific genome to locate the binding sites of the RBP of interest, as shown in **Figure 5B**.

1.6 *In vitro* techniques to study RNA-protein interactions

Despite the strong advantage in detecting RNA-protein interactions *in vivo*, *in vitro* approaches are still preferred to study intrinsic binding properties of RBPs, such as binding site affinity and motif identification. This preference is primarily due to complexity of studying RNA-protein interactions in the cell, which complicates the interpretation of data produced by *in vivo* techniques. Therefore, various *in vitro* techniques have evolved to complement the *in vivo* approaches for studying RNA-protein interactions.

Systematic evolution of ligands by exponential enrichment (SELEX) is one of the most common techniques used to identify the binding motif of an RBP (Marchese et al., 2016; Cook et al., 2014; Ellington & Szostak, 1990). This protocol uses randomized oligo sequences mixed with RBP to detect binding preference. Interacting RNA oligos are then reverse transcribed and amplified. The amplified oligos are used to transcribe a new pool of RNAs and repeat the process several times. This repetition results in the enrichment of oligos with high affinity for binding to the RBP of interest. The identity of these oligos are revealed via cloning, followed by sequencing. To increase the throughput of the detection, several techniques with principles similar to SELEX have since been developed, such as SEQRS and RNA-compete, both of which employ high-throughput sequencing or microarray for the detection of the interacting oligos (Campbell et al., 2012; Ray et al., 2009). Further improvement of *in vitro* RNA-protein interaction approaches was achieved by incorporating a quantitative aspect: not only detecting motif of a specific RBP of interest but also measuring the affinity of the interactions. RNA-Map is the first technique that incorporates this quantitative aspect in a high-throughput platform to study RNA-protein interactions (Buenrostro et al., 2014).

Thus far, all *in vitro* technologies mentioned above focus on interrogating the interaction between RBP and single-stranded RNA sequences. However, secondary structure is also important for influencing RNA-protein interactions and has been shown to play important roles in regulating cellular processes. Interest in these roles have spurred the development of RNA Bind n-Seq, which uses a pool of longer RNA oligos (~40 nt) to study RNA-protein interaction (Lambert et al., 2014). This technique allows the formation of secondary structure that is possibly involved in modulating the RNA-protein interactions. Using a different focus, RNA-MITOMI (mechanically induced trapping of molecular interactions) was developed to specifically identify the interactions between RBPs and RNA secondary structures (Martin et al., 2012). Together, the improvements of *in vitro* approaches to study RNA-protein interaction provide a new layer of information about RBP binding behaviors, which increases the relevance of the output to complement the interpretation of *in vivo* data.

2. AIMS OF THE PROJECT

Despite all the advancements, *in vitro* techniques still use rather short RNAs to study RNA-protein interactions. This artificial length of RNA might undermine the effect of additional factors that come with longer RNA context, namely RNA secondary structure. In addition, currently the *in vitro* approaches utilize different principle in their protocol in comparison to the *in vivo* techniques to capture RNA-protein interaction. Therefore, it is difficult to directly compare their data with *in vivo* interaction data. This problem motivated us to develop *in vitro* iCLIP, an adaptation of the iCLIP protocol for *in vitro* interaction study that follows the same principle in capturing RNA-protein interactions. By using a minimal system consists of U2AF65 and longer RNAs, we would like to use *in vitro* iCLIP to investigate the intrinsic binding behavior of U2AF65 in the absence of other RBPs. Comparing the intrinsic binding *in vitro* to the *in vivo* binding, we aimed to systematically capture the comprehensive U2AF65 regulatory events in a transcript-wide manner. Based on these data, we would then performed a screening to identify possible RBP regulators that shape the functional U2AF65 binding *in vivo*.

3. MATERIALS AND METHODS

3.1 Materials

3.1.1 Buffers

E. coli lysis Buffer:
50 mM Tris, pH 7.5
1 M NaCl
1 mM DTT
10% Glycerol
0.1% TritonX

E. coli wash buffer:
50 mM Tris, pH 7.5
500 mM NaCl
5% (v/v) glycerol
20 mM imidazole.
1 mM DTT

Elution Buffer:
50 mM Tris, pH 7.5
500 mM NaCl
5% glycerol
500 mM imidazole

Binding Buffer
10 mM Hepes, pH 7.2
3 mM $MgCl_2$
3% glycerol
1 mM DTT

PNK Buffer
20 mM Tris-HCl, pH 7.4
10 mM $MgCl_2$
0.2% Tween-20

5x PNK pH 6.5 Buffer:
350 mM Tris-HCl, pH 6.5
50 mM $MgCl_2$
5 mM DTT

Lysis Buffer/Wash buffer
50 mM Tris-HCl, pH 7.4
100 mM NaCl
1% Igepal CA-630
0.1% SDS
0.5% sodium deoxycholate

High-salt Wash
50 mM Tris-HCl, pH 7.4
1 M NaCl
1 mM EDTA
1% Igepal CA-630
0.1% SDS
0.5% sodium deoxycholate

4x Ligation Buffer
200 mM Tris-HCl, pH 7.8
40 mM $MgCl_2$
4 mM DTT

PK Buffer
100 mM Tris-HCl, pH 7.4
50 mM NaCl
10 mM EDTA

PK Buffer + 7 M Urea
100 mM Tris-HCl, pH 7.4
50 mM NaCl
10 mM EDTA
7 M urea

Diffusion Buffer
0.5 M $C_2H_7NO_2$
10 mM $C_4H_6MgO_4$
1 mM EDTA
0.1 % SDS

Cell lysis buffer
50 mM Tris-HCl, pH 8.0
150 mM NaCl
1% Igepal CA-630

PBST, pH 7.4
3.2 mM Na_2HPO_4
0.5 mM KH_2PO_4
1.3 mM KCl
135 mM NaCl
0.05% Tween-20

MST buffer
50 mM Tris-HCl, pH 7.4
150 mM NaCl
10 mM $MgCl_2$
0.05% Tween-20

3.1.2 Reagents and disposables

Name	Cat. no.	Company
NuPage 4-12% BT Gel	NP0322BOX	Thermo Fisher Scientific
Phenol:chloroform:isoamyl alcohol	P3803-400mL	Sigma-Aldrich Chemie GmbH
Protein G Dynabeads	10004D	Life technologies
Glycoblue	AM9516	Life technologies
TBE-Urea Sample Buffer	LC6876	Life technologies
SYBR Gold	S11494	Life technologies
Protease inhibitor cocktail	P8340-5mL	Sigma-Aldrich Chemie GmbH
Phase lock gel heavy tubes	713-2536	VWR International GmbH
Nitrocellulose membrane	10600002	VWR International GmbH
Prestained protein marker	P7712 S	New England Biolabs
Deoxynucleotide Solutions	N0447 L	New England Biolabs
Spin-X centrifuge tube filters	CLS8161-100EA	Corning
RNase Inhibitor	N2615	Promega GmbH
6% TBE-urea gel	EC6865BOX	Life technologies
Whatman glass microfiber filters	WHA1823010	Sigma-Aldrich Chemie GmbH
Ethanol absolute	32205-2.5L-D	Sigma-Aldrich Chemie GmbH
Low molecular weight DNA ladder	N3233L	New England Biolabs
Mini clarification spin columns	GEN-MSF500	Serva
ATP, [γ-32P]	NEG 502A250UC	PerkinElmer
PEG400	202398	Sigma-Aldrich Chemie GmbH
Amersham Protran Premium 0.45 NC nitrocellulose membrane	10600002	VWR International GmbH
10X FastDigest Buffer	B64	Thermo Fisher Scientific
High Sensitivity D1000 ScreenTapes	5067-5584	Agilent Technologies

High Sensitivity D1000 Reagents	5067-5585	Agilent Technologies
MOPS running buffer	NP0001	Invitrogen
NuPAGE Transfer Buffer	NP00061	Invitrogen
NuPAGE Sample Reducing Agent	NP0004	Invitrogen
Lipofectamine RNAiMAX	13778150	Thermo Fisher Scientific
DMEM	11960044	Thermo Fisher Scientific
Penicillin-Streptomycin	15140122	Thermo Fisher Scientific
L-Glutamine	25030081	Thermo Fisher Scientific
Fetal Bovine Serum	10500-056	Thermo Fisher Scientific
Opti-MEM I Reduced-Serum Medium	31985062	Thermo Fisher Scientific
Monolith NT.115 Capillaries	MO-K002	NanoTemper Technologies
Complete Protease Inhibitor, Mini, EDTA-free	4693159001	Sigma-Aldrich Chemie GmbH
IPTG	10725471	Thermo Fisher Scientific
Ni Sepharose	17-5318-01	GE Healthcare
Imidazole	I5513	Sigma-Aldrich Chemie GmbH
Spin-X UF 500 concentrators	431477	Corning
Oligo(dT)$_{18}$	SO132	Thermo Fisher Scientific
Igepal CA-630	I8896	Sigma-Aldrich Chemie GmbH
Human Genomic DNA	G304A	Promega GmbH

3.1.3 Enzymes

Name	**Cat. no.**	**Company**
RNase I	AM2295	Life technologies
T4 Polynucleotide kinase	M0201 L	New England Biolabs
T4 RNA ligase	M0204 L	New England Biolabs
BamH1	FD0055	Life technologies

TURBO DNase	AM2238	Thermo Fisher Scientific
Proteinase K	3115828001	Roche
AccuPrime Supermix 1 enzyme	12342010	Invitrogen
One*Taq* DNA Polymerase	M0480S	New England Biolabs
Taq DNA Polymerase	M0273S	New England Biolabs
HindIII-HF	R3104S	New England Biolabs
NotI-HF	R3189S	New England Biolabs

3.1.4 Kits

Name	**Cat. no.**	**Company**
SuperScript III Reverse Transcriptase	18080044	Thermo Fisher Scientific
CircLigase II ssDNA Ligase	131406	Epicenter
Qubit dsDNA HS (High Sensitivity) Assay Kit	Q32851	Thermo Fisher Scientific
LabChip XT DNA 300 Assay Kit	760601	PerkinElmer
RevertAid First Strand cDNA Synthesis Kit	K1622	Thermo Fisher Scientific
Riboprobe Systems - T7 (*in vitro* Transcrption Kit)	P1440	Promega GmbH
QIAprep Spin Miniprep Kit	27106	Qiagen GmbH
Q5 Site-Directed Mutagenesis Kit	E0554S	New England Biolabs
Luminaris HiGreen Low ROX qPCR MM	13505260	Thermo Fisher Scientific
RNeasy Mini Kit	74106	Qiagen GmbH
QIAquick PCR Purification Kit	28106	Qiagen GmbH
QIAquick Gel Extraction Kit	28706	Qiagen GmbH
Phusion High-Fidelity PCR Kit	M0531 S	New England Biolabs
TOPO XL PCR Cloning Kit	K4700	Life technologies
MinElute PCR Purification Kit	28004	Qiagen GmbH
Pierce BCA Protein Assay Kit	23225	Thermo Fisher Scientific

3.1.5 Antibodies

Name	**Cat. no.**	**Company**
anti-U2AF65	U4758	Sigma-Aldrich Chemie GmbH
anti-β-actin	A5316	Sigma-Aldrich Chemie GmbH
anti-mouse IgG HRP-linked	7076	New England Biolabs

3.1.6 Primers

Primer name	Sequence	Comment
f_mat2a	ATTTCTCTTTTCCAGATAAGATTTG	*MAT2A* insert
r_mat2a	ATATAAAGAGTGTATACCTGAACAA	*MAT2A* insert
f_pcbp2_BamHI	TCATGGATCCGGGCAAGTGCTGTTGCTTTT	*PCBP2* insert
r_pcbp2_NotI	TCATGCGGCCGCGCAGGACACTGCTGAAACGAC	*PCBP2* insert
f_mir7_BamHI	TCATGGATCCGCTGCCCCCATCAATCTAGT	*MIRLET7A2* insert
r_mir7_NotI	TCATGCGGCCGTGTGCCATCTACTCGCCATT	*MIRLET7A2* insert
f_mal1_1_NotI	TCATGCGGCCGCCGAATTCCGGTGATGCGAGT	*MALAT1* insert 1
r_mal1_1_XbaI	TCATTCTAGACCAATATTTGCCCCTCCCCT	*MALAT1* insert 1
f_mal1_2_NotI	TCATGCGGCCGCAGTACAGCACAGTGCAGCTT	*MALAT1* insert 2
r_mal1_2_XbaI	TCATTCTAGAGCAGCGGGATCAGAACAGTA	*MALAT1* insert 2
f_mal1_3_NotI	TCATGCGGCCGTCAGGTCTGTCTGTTCTGTTGG	*MALAT1* insert 3
r_mal1_3_XbaI	TCATTCTAGACAGGGATTTGAACCCCGTCC	*MALAT1* insert 3
f_myc_BamHI	TCATGGATCCAAAGGGGGTGAAAGGGTGCT	*MYC* insert
r_myc_NotI	TCATGCGGCCGCCTGCGTAGTTGTGCTGATGTG	*MYC* insert
f_myl6_BamHI	TCATGGATCCTGACTGTAGCTATACCCTCTGG	*MYL6* insert
r_myl6_NotI	TCATGCGGCCGCAAATTCACACAGGGAAAGGCA	*MYL6* insert
f_ptbp2_BamHI	TCATGGATCCTCAACCGGGAGAGTGTTGTG	*PTBP2* insert
r_ptbp2_NotI	TCATGCGGCCGCTCTTCACACGCTGCACATCT	*PTBP2* insert

f_c4bpb_NotI	TCATGCGGCCGCCCACAGGATCCCTGTCTTTTTA	*C4BPB* insert
r_c4bpb_XbaI	TCATTCTAGATTTATCTTGTCTTGGCGGAGAG	*C4BPB* insert
f_nf1_BamHI	TCATGGATCCTAATTACAGGGCTCGTCCAACT	*NF1* insert
r_nf1_NotI	TCATGCGGCCGCAAAGCATTTGTGAGCTGCAGTA	*NF1* insert
f_papd4_BamHI	TCATGGATCCAATACTTCAGGCTTTGCCACTC	*PAPD4* insert
r_papd4_NotI	TCATGCGGCCGCAAGGTGAGTATATGCCGTGCTT	*PAPD4* insert
fw_PCBP1_qPCR	AAGACTTGACCACGTAACGAG	qPCR primer for *PCBP1* KD check
rev_PCBP1_qPCR	ATGCTTCCTACTTCCTTTCCG	qPCR primer for *PCBP1* KD check
fw_FUBP1_qPCR	ACGGGCTGGAGTTAAAATGG	qPCR primer for *FUBP1* KD check
rev_FUBP1_qPCR	TCTCTGAAACCGCCTTGATC	qPCR primer for *FUBP1* KD check
fw_SNRPA_qPCR	ATCTTGTTCCTCACCAACCTG	qPCR primer for *SNRPA* KD check
rev_SNRPA_qPCR	ACCTCATTGTCAAACTCCACG	qPCR primer for *SNRPA* KD check
fw_RBM41_qPCR	CCAAGCGACTTCTCTCATATC	qPCR primer for *RBM41* KD check
rev_RBM41_qPCR	CATCCAGGACCCACAATTTA	qPCR primer for *RBM41* KD check
qPCR_RBM24_fw1	AAAGACCTTTCGGGATACCTGC	qPCR primer for *RBM24* KD check
qPCR_RBM24_rv1	TGCGTATGCAGCTCCAGTG	qPCR primer for *RBM24* KD check
qPCR_CELF6_fw1	TCCCCTTATCCAGCCCAGAG	qPCR primer for *CELF6* KD check
qPCR_CELF6_rv1	TAGTGGTGCATCCCAGCGTA	qPCR primer for *CELF6* KD check
qPCR_MBNL1_fw1	CCCAGCAAATGCAACTAGCC	qPCR primer for *MBNL1* KD check
qPCR_MBNL1_rv1	ACTGAAAACATTGGCACGGG	qPCR primer for *MBNL1* KD check
qPCR_ELAVL1_fw1	CATTAAGGTGTCGTATGCTCGC	qPCR primer for *ELAVL1* KD check

qPCR_ELAVL1_rv1	GAGCCCGCTGATGTACAAGT	qPCR primer for *ELAVL1* KD check
qPCR_KHDRBS1_fw1	AGAGTCAAGGGGAGTCAGAGT	qPCR primer for *KHDRBS1* KD check
qPCR_KHDRBS1_rv1	AGTGATGGCCTGGTCCCATT	qPCR primer for *KHDRBS1* KD check
qPCR_HNRNPC_fw1	GAACCCGGGAGTAGGAGACT	qPCR primer for *HNRNPC* KD check
qPCR_HNRNPC_rev1	AGCCGAAAATGTAGCTGAAGA	qPCR primer for *HNRNPC* KD check
exon9_exon11_nPTB FWD Set 2	GCAATACAGTCCTGTTGGTTAG	primer to check alt. exon inclusion on *PTBP2*
exon9_exon11_nPTB REV Set 2	GATTGGTTTCCATCAGCCATCT	primer to check alt. exon inclusion on *PTBP2*
fw_CD55_alt_exon	TCAGGTACTACCCGTCTTCTATC	primer to check alt. exon inclusion on *CD55*
rev_CD55_alt_exon	GGAACAGTCTGTATACTTGTGTGTA	primer to check alt. exon inclusion on *CD55*
fw_PCBP2_alt_exon	TTGACCAAGCTGCACCAG	primer to check alt. exon inclusion on *PCBP2*
rev_PCBP2_alt_exon	TCGTTTGGAATGGTGAGTTCAT	primer to check alt. exon inclusion on *PCBP2*
fw_MYL6_alt_exon	CAGCAATGGTTGTATCAACTATGAA	primer to check alt. exon inclusion on *MYL6*
rev_MYL6_alt_exon	GGAAAGTTGCTGAGACAAGAAAG	primer to check alt. exon inclusion on *MYL6*
fw_ptbp2_mut1	ACGCAATGTTGAGATGAAATGCTGTAATTTAC	primer to construct mutant BS3sub
rev_ptbp2_mut1	TGCATTGCATGCATAGGAATTTTAAACTTTG	primer to construct mutant BS3sub
fw_ptbp2_del	TTGTCTTCATTCCCTGTC	primer to construct mutant BS2del
rev_ptbp2_del	TCAGTACATAGGAAATGCAG	primer to construct mutant BS2del
fw_ptbp2_del4	ACGCTGCTTGCTCTTCTC	primer to construct mutant BS1del
rev_ptbp2_del4	GATGAACAGATCATCCATAACG	primer to construct mutant BS1del

fw_ptbp2_subs1	TGTTTACAGTTTACAGTTCCTTGTCTTCATTCCCTGTC	primer to construct mutant BS2sub
rev_ptbp2_subs1	AATGCTGCAAAGGTTAACTATCAGTACATAGGAAATGC	primer to construct mutant BS2sub
rev_vector_minigene	GCAACTAGAAGGCACAGTCG	primer to check alt. exon inclusion on PTBP2 in minigene
RT10_primer_iCLIP	NNGACCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT16_primer_iCLIP	NNTTAANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT19_primer_iCLIP	NNAATANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT20_primer_iCLIP	NNACGCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT30_primer_iCLIP	NNCGATNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT31_primer_iCLIP	NNCGTANNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT32_primer_iCLIP	NNCTCGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT40_primer_iCLIP	NNGGCGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT41_primer_iCLIP	NNGTATNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT48_primer_iCLIP	NNTGTGNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT49_primer_iCLIP	NNTTCTNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation
RT50_primer_iCLIP	NNTTTCNNNAGATCGGAAGAGCGTCGTGgatcCTGAACCGC	primer for reverse transcription during *in vivo* and *in vitro* iCLIP library preparation

3.1.7 siRNAs

Target gene	Source	ID	Sequence	Working conc. (nM)
U2AF65	ON-TARGET Plus SMARTpool siRNA Dharmacon	L-012380-02-0005	CGGUAGGAACAUAGCGUGU	5
			CCAUGCAAGCUGCGGGUCA	
			AGGAGAACCGGCAUCGGAA	
			GCGGCAGCUCAACGAGAAU	
hnRNPC	single siRNA Invitrogen	HNRNPCHSS179305	AAGCAGUAGAGAUGAAGAAUGAUAA	5
PTBP1	ON-TARGET Plus SMARTpool siRNA Dharmacon	L-003528-00	CGUCAAAGGAUUCAAGUUC	5
			GGCACAAGCUGCACGGGAA	
			GAACUUCCAGAACAUAUUC	
			GCAUCACGCUCUCGAAGCA	
PTBP2	ON-TARGET Plus SMARTpool siRNA Dharmacon	L-021323-01	GAGAGGAUCUGACGAACUA	5
			UGACAUGACUUACGUGCAU	
			GGAACUAGCAACCGAGGAA	
			AGAAGAGGAUCUACGAACA	
BRUNOL6	MISSION pre-designed siRNA Sigma	SASI_Hs01_00205595	GGAGUUUGGUGAUGCGGAA	20
MBNL1	MISSION pre-designed siRNA Sigma	SASI_Hs02_00354513	GCAACAACAUCUGCCACAA	20
ELAVL1	MISSION pre-designed siRNA Sigma	SASI_Hs01_00049654	GGCUUGAGGCUCCAGUCAA	20
FUBP1	custom siRNA	-	CUGGAACACCUGAAUCUGU	20
SNRPA	MISSION pre-designed siRNA Sigma	SASI_Hs01_00224233	GAUAUCAUUGCCAAGAUGA	20
KHDRBS1	MISSION pre-designed siRNA Sigma	SASI_Hs01_00219781	CAUAUGAAGAAUAUGGAUA	20

RBM24	MISSION pre-designed siRNA Sigma	SASI_Hs01_00094931	GCAAUAUGUAGCUUGAAUU	20
RBM41	MISSION pre-designed siRNA Sigma	SASI_Hs01_00218259	CAGAAAUCUUGAUCUGGAA	20
PCBP1	MISSION pre-designed siRNA Sigma	SASI_Hs01_00034329	CGGUUAAGAGGAUCCGCGA	20
control	custom siRNA	-	UGGUUUACAUGUCGACUAA	20

3.2 Methods

3.2.1 Preparation of recombinant proteins

6xHis-tagged U2AF65^{RRM12} and hnRNPC1 recombinant constructs were overexpressed in an *E. coli* BL21-CodonPlus(DE3)-RIL strain. Briefly, glycerol stock containing *E. coli* strain carrying plasmid construct was pre-cultured overnight at 37ºC 250 rpm under antibiotic selection (50 µg/ml of chloramphenicol and 100 µg/ml kanamycin). In the following day 2.5 ml the pre-cultured *E. coli* was inoculated into 100 ml fresh LB media and cultured until OD_{600} = 0.8 was reached. A final IPTG concentration of 0.5 µM was added to the culture to induce protein expression. The induced culture was then grown for another 4 hr at 37ºC 250 rpm and harvested by centrifugation at 3000 *xg* for 5 min at 4ºC. The supernatant was discarded and the pellet was stored at -80ºC. For protein extraction, the collected pellet was lysed by adding 4 ml *E.coli* lysis buffer containing Complete Protease Inhibitor followed by 1 hr vortexing at 4ºC. The lysate was then centrifuged at 21000 *xg* for 1 hr 4ºC. The supernatant was collected and 1/10 volume of Ni-sepharose beads were added to affinity purify the recombinant protein from the lysate. The affinity purification was performed for 1 hr at 4ºC with rotation. The beads were then collected by centrifugation and washed 10x with *E. coli* washing buffer. The recombinant protein was eluted by adding 150 µl of elution buffer to the beads and spinning it through proteus clarification column. The flowthrough containing the recombinant protein was kept for the subsequent process. The recombinant protein concentration was quantified with BCA protein assay kit after buffer exchange with binding buffer in Spin-X UF concentrator. The quality of the purification was determined with SDS-PAGE. Additional purification with size exclusion chromatography was performed by Maximilian Bach (master student) to achieve better purity for hnRNPC1 recombinant protein preparation.

Recombinant FLAG-tagged PTBP1 expressed and purified from mammalian cells was obtained from Kelifa Arab (Heidelberg University). Recombinant 6xHis-tagged full-length U2AF65 produced in *E. coli* was obtained from Dr. Hyun-Seo Kang (Technical University Munich). For the co-factor experiments, all GST-tagged recombinant RBPs except for hnRNPC1 and PTBP1 (BRUNOL6, ELAVL1, FUBP1, KHDRBS1, MBNL1, PCBP1, RBM24, RBM41 and SNRPA) were purchased from Abnova as *in vitro* translation products.

3.2.2 Preparation of in vitro transcript mix

In total, 11 different *in vitro* transcripts were used for the *in vitro* iCLIP experiments. The transcripts were chosen to harbor a diverse set of constitutive and alternative exons as well as to show high coverage of U2AF65 *in vivo* iCLIP reads, which facilitated comparative *in vitro – in vivo* analysis. Briefly, the selected regions were amplified with target specific primers from Human Genomic DNA template using Phusion High-Fidelity PCR Kit according to manufacturer's instructions. The amplified products were then inserted into pcDNA3 vector via restriction ligation or pCR2.1 vector via topo cloning by using TOPO XL PCR Cloning Kit according to manufacturer's instructions. All constructs were then transformed into *E. coli* DH5α for replication. Positive clones were confirmed by using Sanger sequencing. Constructs harboring genes for the transcript set were then *in vitro* transcribed by using Riboprobe System-T7 according to the manufacturer's instructions. The *in vitro* transcripts were then treated with TURBO DNase for 15 min at 37^{o}C and purified with RNeasy MinElute Cleanup Kit. Concentrations of each purified transcript were determined with a NanoDrop 2000 system to estimate the required volume for making the stock of equimolar *in vitro* transcript mix. All *in vitro* transcripts were kept in -80^{o}C until subsequent use.

3.2.3 In vitro iCLIP library preparation and sequencing

The *in vitro* iCLIP protocol was developed by modifying the early steps of the standard iCLIP protocol (Huppertz et al. 2014; Sutandy et al. 2016). Briefly, beads were prepared by 2x washes of 40 µl of protein-G Dynabeads per sample with dilution buffer (corresponding to the lysis buffer in the *in vivo* iCLIP protocol). After the second wash, 40 µl of dilution buffer was added to resuspend the beads and then followed by addition of 3 µg of anti-U2AF65 antibody. The beads were rotated at room temperature for 30-60 min. One-time high-salt buffer and 2x dilution buffer washes were applied to wash the beads before proceeding with immunoprecipitation.

In vitro transcripts mix were preheated for 5 min at 70°C to reduce large-scale RNA secondary structures. Titrated concentrations of U2AF65^{RRM12} (150 nM, 250 nM, 450 nM,

750 nM, 1.5 µM, 3 µM, 5 µM, 15 µM) and 2.2 nM *in vitro* transcript mix (eleven transcripts) were used for the K_d measurements. For initial hnRNPC1 titration experiment, 1 µM U2AF65^{RRM12} was mixed with 6.75 nM *in vitro* transcript mix (nine transcripts; excluding *MALAT1* and *MIRLET7A2*) and different concentrations of recombinant hnRNPC1 (200 nM, 500 nM, and 1 µM) in binding buffer. For *in vitro* co-factor experiments, 500 nM U2AF65^{RRM12} was mixed with 6.75 nM *in vitro* transcript mix (nine transcripts) and different concentrations of 11 recombinant RBPs in binding buffer. In addition, 500 nM BSA was added to 500 nM U2AF65^{RRM12} and 6.75 nM *in vitro* transcript mix as a control for the *in vitro* iCLIP co-factor experiments. Moreover, to test the linearity between input material and output of *in vitro* iCLIP experiments, five different dilutions (1x, 2x, 4x, 8x and 16x) of a mixture between 2.5 µM U2AF65^{RRM12} and 6.75 nM *in vitro* transcripts (nine transcripts) were prepared.

All *in vitro* mixtures were incubated for 10 min at 37°C. After the incubation, the mixtures were placed on a parafilm-coated plate on top of an ice plate and UV-irradiated with 5 mJ/cm^2 250 nm UV wavelength (Stratalinker 2400). Since only a minor fraction of the overall interactions (<5%) are expected to be crosslinked during this time, the irradiation should not dramatically shift the binding equilibrium. The irradiated *in vitro* mixtures were pooled back to the tubes, and dilution buffer was added to fill the samples to a volume of 1 ml. To normalize the final *in vitro* iCLIP libraries, 10 µl of the crosslinked mixture containing 250 nM U2AF65^{RRM12} and 6 nM *NUP133 in vitro* transcript was spiked in to each sample. Partial RNase digestion was performed by adding 10 µl of 1:1500 diluted RNase I to each sample. In addition, 2 µl of TURBO DNase was added to each sample to avoid DNA contamination. The sample mixtures were incubated for 3 min at 37°C, added to the prepared beads, and incubated for 2 h at 4°C. The beads were washed twice with high-salt buffer and twice with wash buffer.

Henceforth, we followed the steps of the standard iCLIP protocol. Briefly, 3' end RNA dephosphorylation was performed by resuspending the beads in 20 µl of mixture containing 4 µl of 5x PNK buffer, 0.5 µl of PNK, 0.5 µl RNasin Ribonuclease Inhibitor, and 15 µl water, followed by incubation for 20 min at 37°C. The beads were washed once with wash buffer, once with high-salt buffer and twice with wash buffer.

For linker ligation, pre-adenylated L3 linker (5′-App-AGATCGGAAGAGCGGTTCAG-dideoxycytidine-3') was ligated by resuspending the beads in the ligation mixture containing 5 µl of 4x ligation buffer, 1 µl T4 RNA ligase, 0.5 µl RNasin, 1.5 µl pre-adenylated L3 linker (20 µM), 4 µl PEG400, and 8 µl water. The samples were incubated at 16°C overnight. The next day, the samples were washed twice with high-salt buffer and twice with wash buffer.

The interacting RNAs were radioactively labeled by resuspending the beads in hot PNK mix (0.2 µl PNK, 0.4 µl 10x PNK buffer, 0.4 µl ^{32}P-γ-ATP, and 3 µl water). The beads were incubated at 1,100 rpm for 5 min at 37°C. The supernatants were removed and the beads were boiled in 20 µl 1x NuPAGE loading buffer (Invitrogen) for 5 min at 70°C. The boiled beads were placed on a magnetic rack. The supernatants were then loaded into the 4-12% NuPAGE Bis-Tris gel and run in 1x MOPS buffer for 50 min at 180 V. RNA-protein complexes from the gel were transferred to a nitrocellulose membrane for 1 h at 30 V.

To extract the interacting RNAs, the membrane was cut into pieces and digested with 10 µl proteinase K in 200 µl PK buffer (100 mM Tris-HCl pH 7.4, 50 mM NaCl, 10 mM EDTA) for 20 min at 37°C. Another 200 µl PK buffer containing 7 M urea were added for further 20 min incubation at 37°C. The RNA-containing mixtures were transferred to Phase Lock Gel Heavy tubes and mixed with 400 µl phenol/chloroform by shaking with 1,100 rpm for 5 min at 30°C. RNAs were extracted by centrifugation for 5 min at 16,000 xg to separate the phases followed by transferring the top aqueous phase containing RNAs to new tubes. The samples were then mixed with 0.75 µl glycoblue, 40 µl 3 M sodium acetate pH 5.5 and 1 ml ethanol absolute, and incubated overnight at -20°C. To precipitate the RNAs, the samples were centrifuged with 21,000 xg for 20 min at 4°C, washed with 80% ethanol, and resuspended in 5 µl water.

cDNA synthesis was performed by adding 1 µl dNTP mix and 1 µl RT primers containing different barcode sequences to each sample (listed in **Materials** section), and incubating them for 5 min at 70°C. The reaction was started by adding RT mixture (4 µl 5x RT buffer, 1 µl 0.1 M DTT, 0.5 µl RNasin, 0.5 µl Superscript III, 7 µl water) to the samples and incubating them for 5 min at 25°C, 20 min at 42°C, 40 min at 50°C, 5 min at 80°C, and hold at 4°C. To hydrolyze the hot RNA templates, 1.65 µl 1 M NaOH was added, followed

by 20 min incubation at 98°C. After the incubation, 20 µl 1 M Hepes-NaOH was added to neutralize the samples' pH. The cDNA libraries were mixed with 0.75 µl glycoblue, 40 µl 3 M sodium acetate pH 5.5 and 1 ml ethanol absolute, and incubated overnight at -20°C. The next day, the cDNA libraries were precipitated by spinning the samples with 21,000 xg for 20 min at 4°C, washing with 80% ethanol, and resuspension in 6 µl water.

The cDNA libraries were mixed with 6 µl 2x TBE-urea loading buffer, heated for 5 min at 80°C, and then loaded and run in a 6% TBE-urea gel for 40 min at 180 V. DNA low molecular weight size marker was used as the ladder. The libraries were size-selected by cutting out the gel within the range of 80-100 nt based on the ladder. Each piece of the gel was then crushed into smaller pieces and mixed with 400 µl diffusion buffer. The mixtures were incubated for 30 min at 50°C, and moved to a Costar SpinX column prepared with two 1 cm glass pre-filters. To extract the cDNA libraries, the mixtures were spun at 16,000 xg for 5 min, and the eluates were added together with 400 µl phenol/chloroform into a Phase Lock Gel Heavy tube. The samples were incubated for 5 min at 30°C, and spun at 16,000 xg for 5 min to separate the phases. The aqueous top layers containing the cDNA libraries were moved to new tubes, mixed with 1 µl glycoblue, 40 µl 3 M sodium acetate pH 5.5 and 1 ml ethanol absolute, and then stored at -20°C overnight.

For circularization, the cDNA libraries were centrifuged with 21,000 xg for 20 min at 4°C, washed with 80% ethanol, and resuspended in 8 µl of ligation mixture (0.8 µl 10x CircLigase buffer II, 0.4 µl 50 mM $MnCl_2$, 0.3 µl CircLigase II, and 6.5 µl water). The cDNA libraries were transferred into PCR tubes and incubated for 1 h at 60°C. To re-linearize the cDNA libraries, 30 µl oligo annealing mix containing 3 µl FastDigest buffer, 1 µl 10 µM cut_oligo (5′-GTTCAGGATCCACGACGACGACGCTCTTCaaaa-3'), and 26 µl water were added. The annealing program was performed by running the samples in successive cycles of 20 seconds from 95°C to 25°C with decreasing the temperature by 1°C in each cycle. After the end of the program, 2 µl of BamHI was added to each sample followed by incubation for 30 min at 37°C and heat inactivation for 5 min at 80°C. The samples were mixed with 350 µl TE, 0.75 µl glycoblue, 40 µl 3 M sodium acetate pH 5.5 and 1 ml ethanol absolute, and then precipitated overnight at -20°C. The next day, the cDNA libraries were extracted by spinning the samples with 21,000 x*g* for 20 min at 4°C, washing with 80% ethanol, and resuspension in 20 µl water.

The libraries were amplified by mixing the cDNA libraries in a PCR reaction containing 0.5 µM P3/P5 Solexa primers mix, and 1x Accuprime Supermix 1 enzyme. The PCR mixes were run with a program comprising a 2 min denaturation step at 94°C, 17-25 cycles of 15 seconds at 94°C, 30 seconds at 65°C and 30 seconds at 68°C, and a final elongation step for 3 min at 68°C. Several pre-PCR steps were performed to estimate the minimal number of cycles that is necessary to amplify the libraries. The amplified libraries were pooled together by purification with the MinElute PCR purification kit. The purified library was size-selected with LabChip XT DNA 300 kit to remove residual P3/P5 Solexa primers from the library. The final libraries were quantified with the Qubit dsDNA HS assay kit and sequenced as single-end reads on an Illumina MiSeq sequencing system.

3.2.4 in vivo iCLIP library preparation and sequencing

In vivo iCLIP libraries were prepared from HeLa cells under *U2AF65* knockdown and wild-type conditions according to the previously published protocols (Huppertz et al. 2014; Sutandy et al. 2016). The HeLa cells were obtained from ATCC (number: CCL-2). For immunoprecipitation, we used 7.5 µg monoclonal anti-U2AF65 antibody produced in mouse per sample. Each iCLIP library was done in triplicates. The libraries were sequenced as single-end reads on an Illumina HiSeq 2500 and NextSeq 500 sequencing system.

3.2.5 Characterization of U2AF65 binding sites

iCLIP sequencing reads were filtered and mapped to the human genome (hg19/GRCh37). Peak calling was performed on combined normalized *in vitro* and *in vivo* iCLIP data by iteratively identifying 9-nt windows with the highest cumulative signal and sufficient enrichment over a region-wise uniform background distribution. This procedure yielded a total of 795 binding sites. To compare the RNA sequence composition at U2AF65 binding sites, we counted all 4-mers as well as the occurrence of pyrimidine-rich motifs within the 9-nt peak region. RNAplfold (Bernhart et al. 2006) was used to compute local RNA sequence accessibility. Moreover, we assigned binding sites to three different transcript regions: binding sites within the first 40 nt of an intron or between the start of the Py tract and

the 3' splice site were defined as ‘associated with the 5′ or 3′ splice site’, respectively, while the remaining intron body is referred to as ‘intronic’. This analysis was performed by Dr. Stefanie Ebersberger. Accessibility scoring of U2AF65 binding sites was performed by Dr. Jörg Fallmann.

3.2.6 Model-based estimation of in vitro K_d values

The binding of U2AF65^{RRM12} to the binding sites in the 11 *in vitro* transcripts was modeled using a reversible and monomeric binding model. By assuming steady state, we expressed the concentration of bound U2AF65^{RRM12} on binding site *i* as a function of the dissociation constant K_{di}, and the U2AF65^{RRM12} and binding site concentrations, respectively:

$$[U2AF65{:}Site_i] = \frac{[Site_i]_{total} \cdot [U2AF65]}{K_{di} + [U2AF65]}$$

Model and experimental data were compared by assuming that the *in vitro* iCLIP signal is proportional to the complex concentration, the experimental error ($e^{\sigma Z_i}$; Z_i being an independent normal random variable), the binding site-specific ‘scaling factor’ (SF_i) and an experiment-specific normalization factor (N):

$$Signal_i = SF_i \cdot N \cdot [U2AF65{:}Site_i] \cdot e^{\sigma Z_i}$$

The unknown parameters SF_i, N, k_{di}, σ were estimated by separately fitting the simulated *in vitro* iCLIP signal to four replicate *in vitro* iCLIP titration experiments. For all replicates and experimental conditions, we assumed the same values for K_{di}, SF_i and $[Site_i]_{total}$, and a relative (log-constant) error σ, whereas N differs between experiments and replicates. Since U2AF65^{RRM12} was present in excess over its target RNAs under *in vitro* conditions, we neglected that the protein may be limiting and therefore set [U2AF65] to be the total U2AF65^{RRM12} concentration in the test tube.

Parameter uncertainties were assessed using the profile likelihood approach (Raue et al. 2009). To this end, each parameter was systematically perturbed around its best-fit value and fixed to this perturbed value, while allowing all remaining parameters to change when refitting the model to the data. This approach yields a two-dimensional profile for each

parameter, the profile likelihood, in which the goodness-of-fit is shown as a function of the fixed parameter value. Finally, a profile likelihood-based confidence interval was calculated for each parameter using the likelihood ratio test at a 95% confidence level (α=0.05, degrees of freedom=1). This modeling analysis was performed by Dr. Lu Huang.

3.2.7 Model-based analysis of in vivo regulatory hotspots

We employed our binding model to systematically identify differences between *in vitro* and *in vivo* binding landscapes. To this end, we searched for the best overlap by fitting the *in vitro* model to the *in vivo* iCLIP landscape. Some biophysical parameters such as k_{di} and SF_i were assumed to be the same *in vitro* and *in vivo*. The unknown concentration of U2AF65, the concentrations of the 29 introns in the 11 transcripts, as well as the *in vivo* experimental error and the *in vivo* normalization factor were estimated by fitting. In contrast to the *in vitro* model fitting, we did not assume the free pool of U2AF65 to be present in excess over the transcripts, and hence allowed for protein sequestration effects between the binding sites.

In order to identify regulatory hotspots at which U2AF65 binding is modulated *in vivo*, we tested at which binding sites the 'expected *in vivo* signal' given by the model fit differs from the *in vivo* measurement. We quantified this difference for binding site *i* and normalized it to the experimental variation to obtain a z-score:

$$z_i = \frac{\ln(Signal_{i,invivo}) - \ln(Signal_{i,model})}{\sigma_{invivo}}$$

Here, σ_{invivo} is the relative error estimated as the standard deviation of the three *in vivo* iCLIP replicates. Binding sites are called as regulated *in vivo* if the difference between model fit and experiment is bigger than the experimental variation ($|z_i| > 1$). The sign of the z-score indicates whether a binding site shows a higher or lower binding affinity *in vivo* when compared to the *in vitro* situation ($z > 1$ and $z < -1$, respectively). This modeling analysis was performed by Dr. Lu Huang.

3.2.8 K_d measurements by MST and ITC

For microscale thermophoresis (MST) experiment, RNA oligonucleotides were selected based on the *in vitro* iCLIP binding landscape. The selected RNA oligonucleotides contain an isolated U2AF65 binding site plus a few nucleotides upstream and downstream of the corresponding site (**Table 2**). 5′-Cy5-labeled selected RNA oligos were chemically synthesized from IDT. Briefly, 5′-Cy5-labeled RNA oligos were mixed to obtain a final reaction containing 150 nM RNA and titrated concentrations of recombinant $U2AF65^{RRM12}$ in MST buffer. Each independent mixture was loaded into an MST capillary. The K_d measurements were then performed with Monolith NT.115 (NanoTemper Technologies) at room temperature according to the manufacturer's instructions and fitted with a Hill equation (Goutelle et al. 2008). For each RNA oligonucleotide, the measurements were done in triplicate.

Isothermal titration calorimetry (ITC) was performed using MicroCal PEAQ-ITC (Microcal) at 25°C. Briefly, 300 µl of 20 µM $U2AF65^{RRM12}$ protein sample (20 mM sodium phosphate, pH 6.5, 50 mM NaCl) in the ITC cell was titrated with 50 µl of 100/100/150/200 µM OR1, 200/200/200 µM OR2, 200/400 µM OR4, 250/300 µM OR5, 200/250 µM OR6, 150/200 µM OR7, and 115/115/130 µM OR8 RNA (IBA) in the same buffer. The data was further analyzed using Origin v5.0 from Microcal. ITC measurement was performed by Dr. Hyun-Seo Kang.

3.2.9 Random Forests analysis

Random Forests (Breiman 2001) was used to classify binding sites into cleared *in vivo* (z-score < -1) or stabilized *in vivo* (z-score > 1). Each binding site was characterized by three types of features in a 99-nt window around U2AF65 binding sites, comprising *k*-mers, position-specific scoring matrices (PSSMs) for 120 unique RBPs (Ray et al. 2013), and positional information, such as splice site strength or distance to the next downstream AG. To identify putative U2AF65 regulators, we considered the top 100 features ranked by importance which were collapsed into 12 regulatory groups. This analysis was performed by Dr. Stefanie Ebersberger.

3.2.10 Analysis of in vitro iCLIP co-factor assays

To facilitate direct comparisons, the reads from each *in vitro* iCLIP co-factor replicate were downsampled, normalized to the spike-in control and converted to 'signal-over-background'. For **Figure 17 & 18**, regulatory categories were assigned according to (i) their model-based *in vivo* regulation (based on the comparison of *in vivo* and *in vitro* U2AF65 binding landscapes) and (ii) *in silico* predictions of associated RBP binding sites. Each set was tested against the control group of U2AF65 binding sites without an associated RBP binding site. To validate the hnRNPC-mediated regulation, we compared our results to previously published *in vivo* U2AF65 iCLIP data from *HNRNPC* knockdown HeLa cells (Zarnack et al. 2013). For **Figure 21**, we used all U2AF65 binding sites within 600 nt upstream of the 3′ splice site (with the exception of *MYL6* exon 6 which shows only 300 nt of preceding intron). This analysis was performed by Dr. Stefanie Ebersberger.

3.2.11 Knockdown of RBPs

HeLa cells were grown in 6-well plate until they reached about 25% confluence. For all RBPs knockdowns, siRNAs were transfected with Lipofectamine RNAiMax reagent according to the manufacturer's instructions. All siRNAs are listed in **Materials** section. The cells were grown for 48 h post transfection and then harvested by scrapping and centrifugation for 3 min at 2000 *xg*. The cell pellets were stored at -80°C for subsequent experiments.

For partial *U2AF65* knockdown, confirmation of the knockdown efficiency was done with Western blot. For the detection, we used a monoclonal mouse anti-U2AF65 antibody and a monoclonal mouse anti-β-actin antibody as primary antibodies, and anti-mouse IgG HRP-linked antibody as secondary antibody.

For knockdowns of all other RBPs in the context of the *in vivo* alternative splicing quantifications, knockdown confirmations were done by measuring mRNA level with Luminaris HiGreen Low ROX qPCR Master Mix in a ViiA 7 Real-time PCR system (Thermo Fisher) according to the manufacturer's instructions. All primers that were used for the measurements are listed in **Materials** section.

3.2.12 Western blot

Cell pellets were lysed with appropriate amount of cell lysis buffer containing Protease Inhibitor Cocktail (~150µl for $4x10^6$ cells). The lysates were then centrifuged for 10 min at 21000 *xg* 4ºC to remove the cell debris. The supernatants were collected in new tubes for the western blot analysis. Total protein content of the lysates were estimated by using Pierce BCA Protein Assay Kit. Briefly, 30 µg of the total protein lysates were mixed with NuPAGE LDS Sample buffer (4x) containing reducing agent and boiled at 80ºC for 10 min. The boiled samples were then loaded and run in 4-12% NuPAGE Bis-Tris protein gels for 50 min at 180V in 1x MOPS running buffer. Following the run, the samples were transferred into Amersham Protran Premium 0.45 NC nitrocellulose membrane in 1x NuPAGE transfer buffer containing 10% methanol for 1hr at 30V. The membrane was blocked with blocking buffer for 30 min and washed 3x with PBST. Appropriate dilution of primary antibody in PBST containing 1% BSA was used to probe specific protein target for overnight at 4ºC. The antibodies used for western blot analysis are listed in the Materials section. In the next day, the membrane was washed 3x with PBST and incubated with appropriate dilution of HRP-conjugated secondary antibody in 5% milk PBST for 45 min at room temperature. The membrane was then washed 3x with PBST and incubated with 750 ml of appropriate chemiluminescence substrate for 1 min. The luminescence signal from the blot was then detected with ChemiDoc station (BioRad) and further analyzed with ImageJ.

3.2.13 Minigene reporter assays

All minigene reporters were constructed by using pCDNA5 backbone via ligation of a 2,727 bp insert containing exon 9-11 of *PTBP2* (Chr1, 96804170 - 96806896 nt). Mutations that were introduced to the different constructs are listed in **Table 3**. The mutant constructs were generated by using the Q5 Site-Directed Mutagenesis Kit according to the manufacturer's instructions. All primers used for the minigene construction are listed on **Materials** section. The *FUBP1* knockdown and the *PTBP1/2* double knockdowns were performed for 48 h as described above. The media were discarded and the cells were further transfected with 2 µg of different minigene constructs (wild-type and mutant variants). The cells were harvested on the next day and the total RNA was extracted with RNeasy Plus Mini

Kit. cDNAs were synthesized with Revert Aid First Strand cDNA Synthesis by using oligo(dT)$_{18}$ primer. The resulting cDNAs were amplified with up to 25 cycles with One Taq polymerase, and the PCR products were visualized in 2200 Tape station system with D1000 DNA screen tape kit to obtain the molar ratio of each splicing product. All primers used in these experiments are listed in **Materials** section. The relative inclusion ('percent spliced in', PSI) in each sample was calculated with the following formula:

$$PSI = \frac{molar\ conc.of\ inclusion\ product}{molar\ conc.of\ inclusion\ product + \ molar\ conc.of\ skipping\ product}$$

3.2.14 In vivo splicing assays

Splicing assays were done by monitoring inclusion of four different alternative exons from *PTBP2, MYL6, CD55,* and *PCBP2* via RT-PCR under control conditions and knockdowns of 11 different RBPs (*CELF6, ELAVL1, FUBP1, HNRNPC, KHDRBS1, MBNL1, PCBP1, PTBP1, RBM24, RBM41* and *SNRPA*). The total RNA was extracted 48 h post-transfection with RNeasy Plus Mini Kit, and the cDNAs were synthesized with Revert Aid First Strand cDNA Synthesis by using oligo(dT)$_{18}$ primer. The resulting cDNAs were amplified with up to 35 cycles with One Taq polymerase, and the PCR products were visualized in 2200 Tape station system. PSI values for each sample were calculated as described above. All primers used in these experiments are listed in **Materials** section.

4. RESULTS

4.1 Establishment of the *in vitro* iCLIP protocol

We designed *in vitro* iCLIP to capture direct RNA-protein interactions in an isolated system, allowing us to investigate the intrinsic binding behavior. The *in vitro* iCLIP protocol was developed based on the established iCLIP protocol (König et al., 2010; Sutandy et al., 2016), which is referred to as *in vivo* iCLIP in this report (**Figure 6**). Instead of UV irradiating living cells, we crosslinked the RNA-protein complexes in an *in vitro* mix containing predefined components. The crosslinked complexes were then immunoprecipitated with specific antibodies targeting the protein to purify them from the *in vitro* mix. To extract the interacting RNAs, proteinase K digestion was applied to the immunoprecipitated complexes. As the *in vivo* iCLIP protocol, *in vitro* iCLIP uses small peptide residue attached to the RNA as a binding footprint that will induce a truncation during subsequent cDNA synthesis. The truncated cDNAs were then circularized and amplified to prepare the *in vitro* iCLIP library for high-throughput sequencing. Mapping the sequencing reads to a genome of origin of the RNAs, will then show the binding locations of the immunoprecipitated protein (**Figure 6**).

The *in vitro* mix consists of recombinant proteins and RNAs in binding buffer. In this project, for the recombinant proteins, we used a short construct of U2AF65 ($U2AF65^{RRM12}$) that contained two RRM domains. This construct has been widely used for *in vitro* studies and shown to be the minimal construct required to largely mimic endogenous U2AF65 binding behavior (Mackereth et al., 2011). For the RNAs, we used *in vitro* transcribed RNAs as our *in vitro* transcript mix that contained 11 different transcripts (unless stated otherwise), representing part of different human genes (**Table 1**). The transcripts were chosen to cover different contexts of splicing regulation.

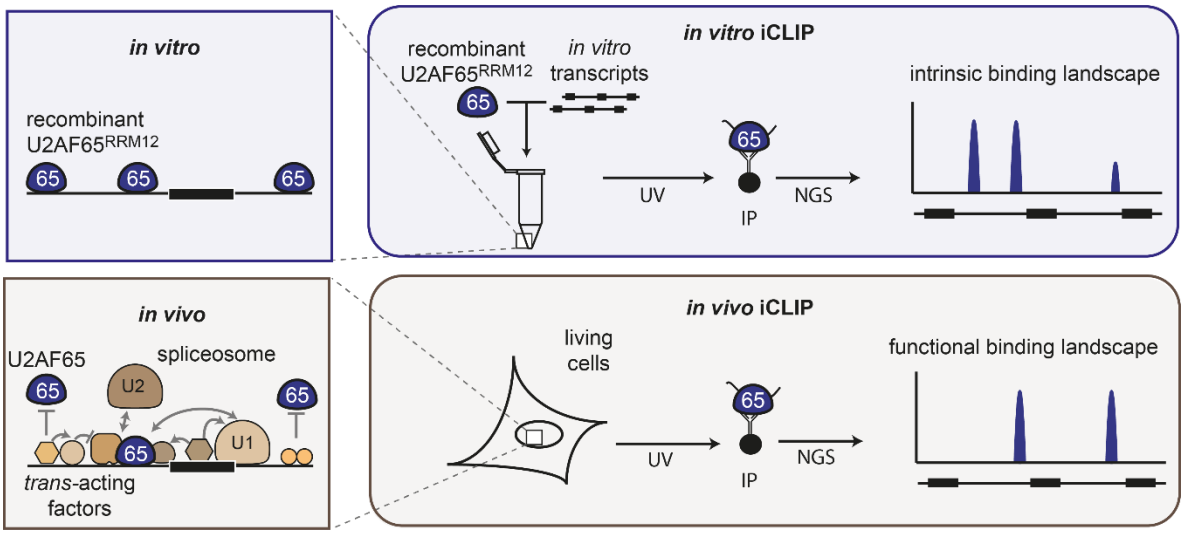


Figure 6. *In vitro* iCLIP protocol. Schematic comparison of *in vitro* and *in vivo* iCLIP. Unlike *in vivo* iCLIP which identifies RNA-protein interactions in the complex cellular environment, *in vitro* iCLIP captures the interactions in a simplified system consisting of naked *in vitro* transcripts and recombinant RBP.

U2AF65 prefers to bind to the Py tract on the RNA. To check if our recombinant U2AF65^{RRM12} preserved this behavior, we performed *in vitro* binding assays by using three chemically synthesized short RNA oligos with different pyrimidine contents (U_9, U_4A_5, and A_9; **Figure 7A**). We crosslinked the *in vitro* mix and visualized the U2AF65^{RRM12} binding strength to the individual oligo in autoradiograph. Here, we showed that U2AF65^{RRM12} bound strongest to the U_9 RNA, and the binding signal decreased in RNA oligos with lower pyrimidine content ($U_9 > U_4A_5 > A_9$; **Figure 7A**).

Table 1. List of *in vitro* transcripts used in *in vitro* iCLIP experiments.

Gene	Chromosome	Coordinate	Strand	Length (kb)
PTBP2	1	97269727:97272451	+	2,725
MAT2A	2	85768258:85769891	+	1,634
MIRLET7A2	11	122016777:122018959	-	2,183
MALAT1	11	65266455:65273637	+	7,182
CD55	1	207512678:207515236 207531777:207533146	+	3,929
PCBP2	12	53860746:53862697	+	1,952
NF1	17	29546069:29550475	+	4,407
PAPBD4	5	78936269:78938704	+	2,436
C4BPB	1	207267907:207270214	+	2,308
MYL6	12	56552950:56555247	+	2,298
MYC	8	128750313:128752817	+	2,505
NUP133	1	229601167:229602503	-	1,336

We further optimized the technical aspects of the *in vitro* iCLIP library preparation, including UV power and partial RNase digestion steps. In iCLIP, the interactions between proteins and different binding sites on the RNA are represented as UV irradiation-induced crosslinking events. Therefore, it is crucial to choose optimal UV power that can efficiently crosslink RNA-protein interactions without oversaturating them and, thus allowing for the discrepancies between different strengths of interactions on the individual binding site. Six different UV powers were applied to crosslink the *in vitro* mixtures. Autoradiograph pictures of the samples showed that the signal from the crosslinked RNAs increases with more UV power used for irradiation in a dose-dependent manner and saturates at approximately 20 mJ/cm^2 (**Figure 7B**). Based on this observation, we chose 5 mJ/cm^2 for the optimal UV power in the standard *in vitro* iCLIP protocol because it produces a crosslinking signal in the linear part of the tested UV crosslinking dynamic range (**Figure 7B**). Despite the changes in overall crosslinking events upon different UV irradiation, the *in vitro* iCLIP binding landscapes are

only slightly affected (**Figure 7C**), indicating that crosslinking powers used to prepare *in vitro* iCLIP libraries did not introduce significant bias to the obtained iCLIP data.

Read length is one of the key factors that defines the quality of high-throughput sequencing data. For most transcriptome analysis, read length greater than 50 bp is considered to be optimal (Chhangawala et al., 2015). In addition, shorter RNA fragments increase the efficiency of library preparation, especially during adaptor ligation and reverse transcription (Ascano et al., 2012). Therefore, to increase the output of the *in vitro* iCLIP library, we partially digested the crosslinked RNAs to enrich for a specific size range that was desired to aid the downstream high-throughput sequencing protocol. We tried five different RNase dilutions and applied them to the *in vitro* mixtures. We labeled the interacting RNAs with γ-P^{32} and extracted them from the complexes via proteinase K digestion. The isolated RNAs were then run in 6% TBE-urea gel to visualize the size distribution, as shown in **Figure 7D**. We chose 1/1500 dilution of RNase for the optimal concentration to obtain RNA fragments with a size range between 50 and 100 nt.

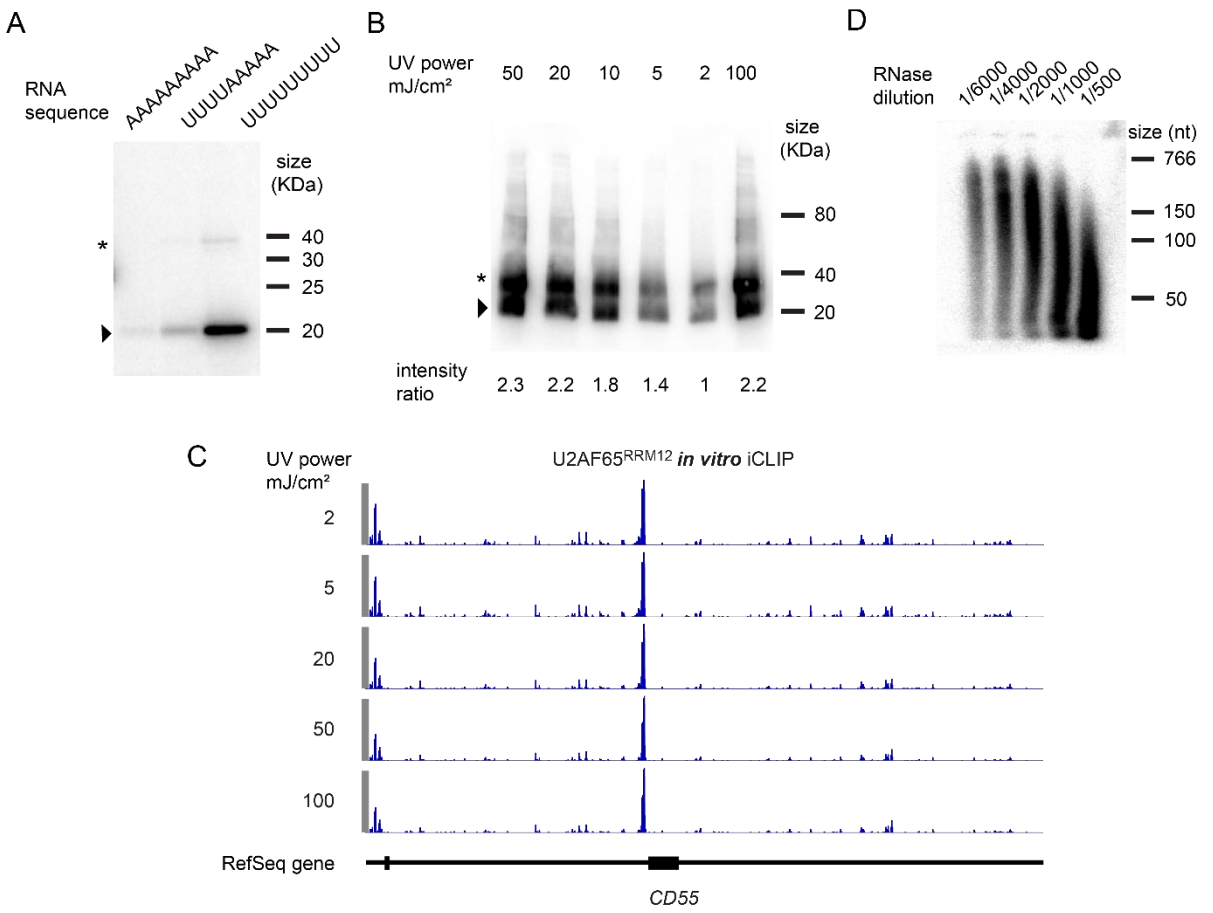


Figure 7. Optimization of *in vitro* iCLIP library preparation. (A) Autoradiograph image of recombinant U2AF65^{RRM12} binding to three RNA oligos with different pyrimidine contents. Arrowhead indicates U2AF65^{RRM12}. Asterisk indicates dimer form of U2AF65^{RRM12}. **(B)** Autoradiograph of U2AF65-RNA complex that were crosslinked with six different UV powers. Relative signal intensities are given below the panel. **(C)** IGV views of U2AF65^{RRM12} *in vitro* iCLIP binding landscape produced with different crosslinking powers at the *CD55* gene. **(D)** Autoradiograph of the extracted RNAs from samples treated with different RNase concentrations. The RNA samples were run in 6% TBE urea gel.

Based on the optimization results, all the following *in vitro* iCLIP experiments were performed with 5 mJ/cm^2 UV power and 1/1500 RNase dilution for the partial RNA digestion.

4.2 U2AF65^{RRM12} resembles full-length U2AF65 binding *in vitro*

With the *in vitro* iCLIP protocol optimized, we performed an initial test of the *in vitro* iCLIP library preparation to first explore the nature of U2AF65^{RRM12} binding behavior *in vitro*. We used 1.5 µM of U2AF65^{RRM12} and *in vitro* transcript mix (11 transcripts) to prepare the *in vitro* iCLIP library. U2AF65^{RRM12} binds to both intronic and exonic parts of the genome *in vitro*, as shown in **Figure 8A**. In total, we found 795 U2AF65 binding sites *in vitro* across 11 different transcripts. To check the reproducibility of the *in vitro* iCLIP protocol, we repeated the experiment in different replicates. **Figure 8B** shows that *in vitro* U2AF65^{RRM12} binding between replicates were well correlated (Pearson $r = 0.904$; p value < 10^{-15}). To further check whether the shortening of the U2AF65 construct would introduce changes to *in vitro* U2AF65 binding behavior, we performed the same experimental setup with a full-length construct of U2AF65 recombinant protein. A comparison between the *in vitro* iCLIP binding landscape of U2AF65^{RRM12} and full-length U2AF65 showed significant positive correlation (Pearson $r = 0.806$, p value < 10^{-15}), indicating no significant changes on the binding behavior existed due to the use of a short U2AF65 construct (U2AF65^{RRM12}; **Figure 8C**).

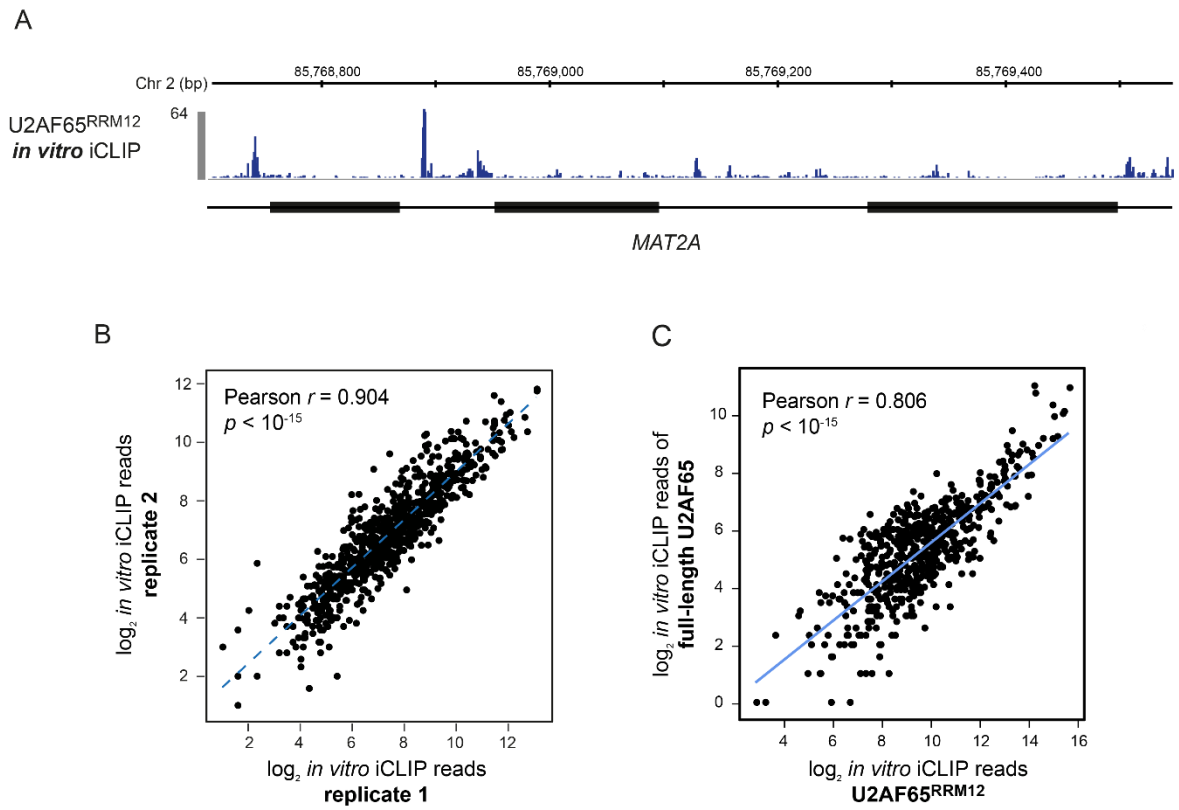


Figure 8. Initial inspection of U2AF65^{RRM12} *in vitro* iCLIP binding landscape.
(A) IGV views of U2AF65^{RRM12} *in vitro* iCLIP binding landscape at the *MAT2A* gene. **(B)** Scatterplot of correlation between U2AF65^{RRM12} *in vitro* iCLIP replicates. **(C)** Scatterplot of correlation between U2AF65^{RRM12} and full-length U2AF65 *in vitro* iCLIP. Pearson *r* and associated *p*-value are indicated in the panel.

4.3 The *in vitro* U2AF65^{RRM12} binding landscape differs from *in vivo* binding

To further determine the correlation between *in vitro* and *in vivo* binding of U2AF65, we compared our U2AF65^{RRM12} *in vitro* iCLIP with previously published U2AF65 *in vivo* iCLIP data (Zarnack et al., 2013; **Figure 9A**). Although several U2AF65 binding sites were preserved between the two landscapes, many of them were differentially regulated. To quantify them, we plotted the distribution of U2AF65 iCLIP reads *in vitro* and *in vivo* across three different transcript regions (**Figure 9B**): 3′ splice site (3′ss), 5′ splice site (5′ss) and intronic region (intron body). We found that U2AF65 binding is highly enriched at 3′ splice sites *in vivo*, supporting the role of U2AF65 in 3′ splice site definition. Intriguingly, the *in vitro* binding of U2AF65 was rather spread across the different transcript regions with minor enrichment at 3′ splice site. Further underlining differences between the *in vitro* and *in vivo* U2AF65 binding landscapes, which were observed by visual inspection. To investigate whether these discrepancies come from binding preference of the recombinant U2AF65^{RRM12} that was used in the *in vitro* experiments, we performed 4-mer analysis of the binding sites from both *in vitro* U2AF65^{RRM12} and *in vivo* U2AF65 iCLIP data. **Figure 9C** shows the significant correlation (Pearson $r = 0.967$, p value $< 10^{-15}$) between the 4-mers frequencies of *in vitro* U2AF65^{RRM12} and *in vivo* U2AF65 iCLIP data, suggesting that the discrepancies between the two landscapes do not rise from intrinsic binding preference of the recombinant U2AF65^{RRM12}.

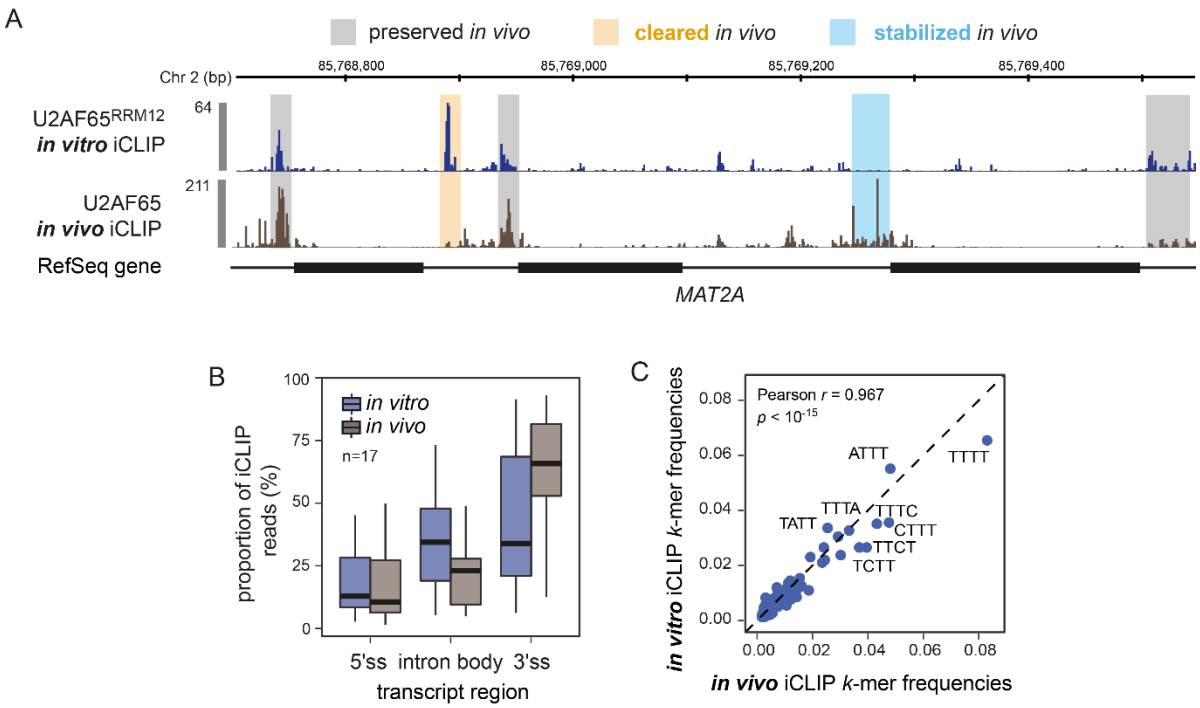


Figure 9. Exploration of U2AF65 *in vivo-in vitro* iCLIP landscapes. (A) IGV view of U2AF65 *in vivo and in vitro* iCLIP data. Peaks that are preserved between *in vivo* and *in vitro* are marked with grey shadow, while peaks that are regulated *in vivo* are marked with orange and blue depending if they were cleared or stabilized, respectively. **(B)** Distribution of *in vivo* and *in vitro* iCLIP reads in three different transcript regions: 5′ splice site (5′ss), intron body, and 3′ splice site (3′ss). **(C)** Scatterplot of 4-mers correlation between *in vivo* and *in vitro* iCLIP of U2AF65. Pearson *r* and associated *p*-value are indicated in the panel.

In conclusion, some discrepancies were observed between the *in vitro* and *in vivo* U2AF65 binding that are not derived from intrinsic binding preference of the recombinant U2AF65^{RRM12}, indicating that U2AF65 binding regulation may occur *in vivo* that is absent in the *in vitro* system.

4.4 Transcript-wide measurement of U2AF65 binding site affinities

In higher eukaryotes, Py tract sequences that precede 3′ splice sites are very degenerate and, therefore, produce variety in U2AF65 binding affinities. In addition, the Py tract is not an exclusive feature of 3′ splice site but rather common in the genome. However, the

enrichment of U2AF65 binding at 3′ splice site is crucial to initiate splicing *in vivo*. Therefore, it is intriguing to determine to what extent the affinity of U2AF65 binding to the Py tract shapes U2AF65 binding distribution *in vivo*. To this end, we performed *in vitro* iCLIP titration experiments to measure transcript-wide U2AF65 binding site affinities in the form of dissociation constants (K_d values). Briefly, increasing concentrations of $U2AF65^{RRM12}$ (0.15–15 µM) were mixed with the *in vitro* transcript mix (11 transcripts) to generate the *in vitro* iCLIP libraries (**Figure 10A**). Using *in vitro* iCLIP reads as a measure for U2AF65-RNA interactions, we fitted *in vitro* iCLIP titration experiments to a mathematical model that generated titration curves and extracted the K_d values of U2AF65-RNA interactions for the individual binding site (**Figure 10B**). The model assumed the interactions occur in a 1:1 equilibrium using mass-action kinetics. Importantly, the model accounts for confounding technical biases such as the sequencing depth of each sample, UV crosslinking efficiencies, PCR amplification biases of RNA sequences and experimental noise. In total, we determined the K_d values for 795 U2AF65 binding sites, which ranged between 0.1 and 1000 µM. The confidence intervals of each K_d value were further determined by using the profile likelihood approach.

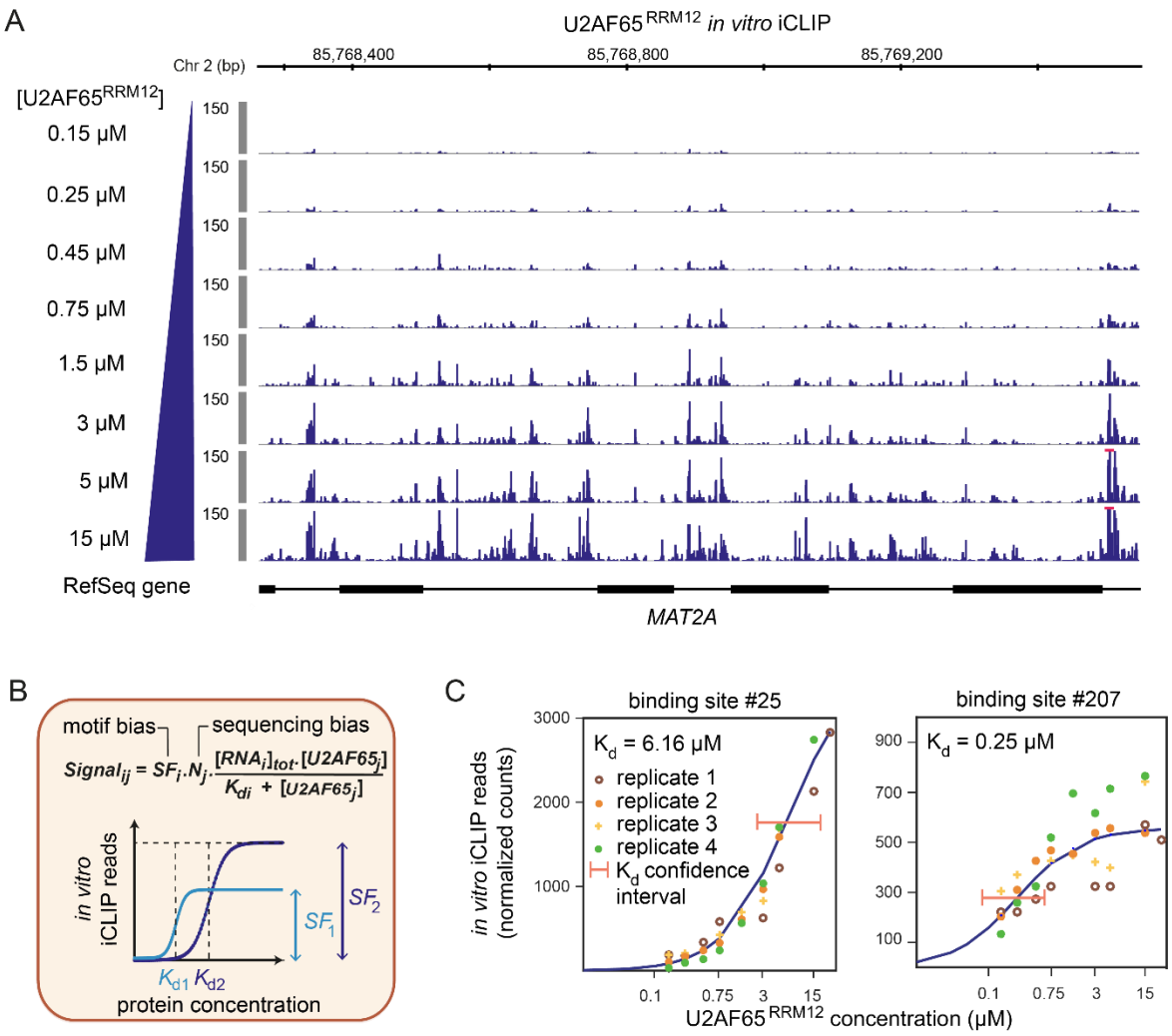


Figure 10. Mathematical modeling of U2AF65 binding site affinities (K_d values) from *in vitro* iCLIP titration assays. (A) IGV view of U2AF65^{RRM12} *in vitro* iCLIP in different concentrations of the recombinant protein at the *MAT2A* gene. **(B)** K_d values were extracted by modeling the read counts from *in vitro* iCLIP as a function of RNA and U2AF65 concentrations. A scaling factor (*SF*) and a normalization factor (*N)* account for motif and sequencing biases, respectively. Schematic titration curves show two binding sites with lower or higher affinity and/or crosslinking efficiency (dark or light blue, respectively). **(C)** Modeled titration curves from individual binding site based on the *in vitro* iCLIP titration experiments.

To further validate the K_d values identified from the *in vitro* iCLIP titration experiments, we performed microscale thermophoresis (MST) and isothermal titration calorimetry (ITC) measurements on eight different chemically synthesized short RNA oligos (one oligo could not be measured on ITC) representing eight different U2AF65 binding sites identified from our iCLIP data (**Table 2**). The correlation plots between K_d values measured with the three different methods are shown in **Figure 11**. The K_d values determined from our *in vitro* iCLIP modeling were significantly correlated with the values measured by both MST (Pearson $r = 0.751$; p value = 0.03) and ITC (Pearson $r = 0.71$; p value = 0.07), indicating that we can reliably use our K_d values as the measure of U2AF65 binding site affinities. Despite the agreement between the three measurements (**Figure 11A-C**), deviations of absolute K_d values occurred between the measurements: 10-fold and 2-fold on average for MST and ITC, respectively, compared with *in vitro* iCLIP (**Figure 11A & B**). These differences were most likely due to the different technical aspects that were used in the measurements, such as the use of short oligos (18-36 nt) in MST and ITC instead of the long RNA sequences used with *in vitro* iCLIP. In addition, the use of Cy5-labeled RNA for MST may have further affected the binding kinetics of the interactions, resulting in larger deviations of absolute K_d values for MST than those measured by *in vitro* iCLIP (**Figure 11A**).

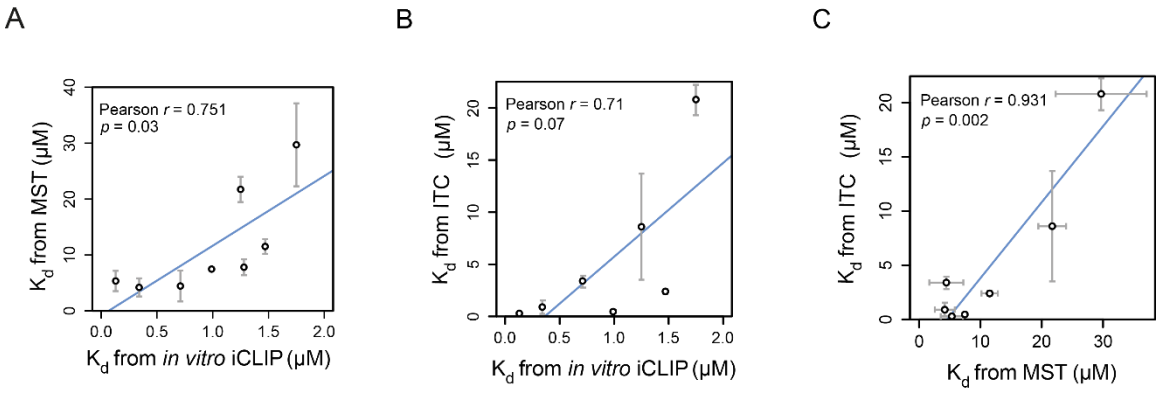


Figure 11. Validation of K_d values measured by *in vitro* iCLIP. Scatter plot of correlation between K_d values measured by **(A)** MST vs *in vitro* iCLIP, **(B)** ITC vs *in vitro* iCLIP, and **(C)** ITC vs MST. Pearson *r* and associated *p*-value are indicated in the panel.

Table 2. RNA oligos that were used in K_d measurement with MST and ITC.

Oligo name	Genomic position			Sequence	Accesibility score
	Chr	start	end		
OR1	8	128,751,896	128,751,932	CUGCAATTTTTTTTTTTTTTATTTTTCATTTCCAGTA	0.24
OR2	11	65,269,035	65,269,053	AAAATGTTTTTTTTCTAAGA	0.08
OR3	1	207,513,317	207,513,336	AGGTCCTTTCTTCTAGTGA	0.06
OR4	1	207,269,096	207,269,113	GGGTGGATTTCTCATGAA	0.01
OR5	12	56,554,916	56,554,938	AGTGGAGAACTTTTCTGCCTCTG	0.06
OR6	2	85,768,883	85,768,904	AGGCTGTTTTAATCTCTTCTAA	0.07
OR7	12	53,861,340	53,861,358	AGGCTGATATTTCTTTGAG	0.08
OR8	1	207,269,352	207,269,371	AAACCTCTTTTCTTTATAAA	0.48

We further investigated the extent of the correlation between the identified K_d values with the pyrimidine content of each individual binding site. We extracted 8-mers motifs from all binding sites and then grouped them based on their pyrimidine content (**Figure 12A**). As expected, binding sites with higher pyrimidine content, such as U_8-mers, showed a higher affinity distribution based on their K_d values. Conversely, binding sites with higher purine (A/G) content showed lower affinities. Intriguingly, we observed that binding sites with similar motifs often have distinct affinities (**Figure 12A**). In collaboration with Jörg Fallman from the Peter Stadler group, we investigated the effect of RNA secondary structures by calculating the probability of each binding site to be accessible for interaction with U2AF65 or buried in RNA structures. Indeed, we saw that RNA secondary structures could introduce significant changes in binding site affinities. For instance, U_8-containing binding sites buried within RNA secondary structures showed substantially lower U2AF65 affinities than binding sites with unstructured U_8-mers that can be freely accessed by U2AF65 (median K_d, 1.75 µM vs. 0.17 µM; **Figure 12A**). Taken together, our data suggested that we can determine reliable K_d values for U2AF65 binding sites and supported the previous findings on U2AF65 binding preference for Py-rich regions.

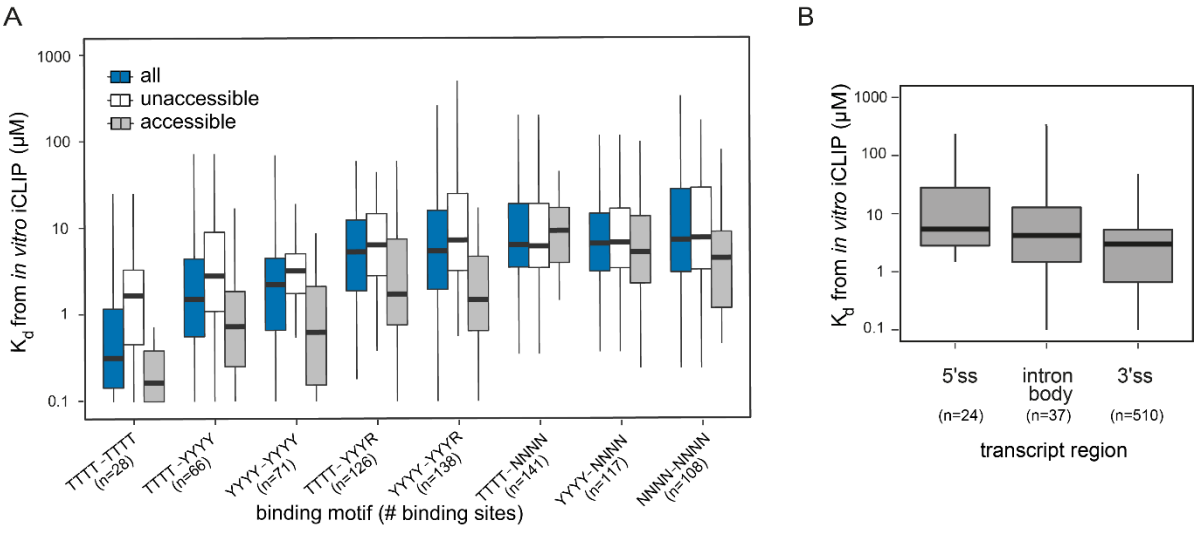


Figure 12. Motif and distribution of binding site affinities measured by *in vitro* iCLIP. (A) Distribution of binding site K_d measured by *in vitro* iCLIP grouped by their 8-mers motif. Unaccessible binding sites are represented as white box, accessible binding site as grey box, and the overall binding sites as blue box. **(B)** Distribution of binding site affinities measured by *in vitro* iCLIP based on their transcript regions (5' splice site, 3′ splice site and intron body).

Ultimately, we checked whether binding site affinities could explain the enrichment of U2AF65 in 3′ splice site *in vivo*. We grouped the binding sites based on their transcript regions and plotted their K_d value distribution (**Figure 12B**). Intriguingly, binding sites at 3′ splice sites showed only slightly higher affinity distribution than other transcript regions. This finding indicates that U2AF65 binding site affinities alone cannot explain the specificity of U2AF65 recruitment to 3′ splice site. Therefore, we speculated that additional regulations from *trans*-acting factors may contribute to shaping the functional U2AF65 binding landscape *in vivo*.

4.5 U2AF65 binding is heavily regulated *in vivo*

The regulation of U2AF65 binding has been widely studied in context of alternative splicing. Here, we found that the importance of such regulation may extend to the general mechanism of U2AF65 recruitment to 3′ splice site. To explore the regulation of U2AF65 binding, we first comprehensively defined the regulated U2AF65 binding sites within our *in vitro* transcript set. We compared the *in vivo* and *in vitro* iCLIP landscapes to identify binding sites that are differentially regulated *in vivo*. To aid in this comparison, the analysis was restricted to the nine *in vitro* transcripts that are derived from the protein-coding genes and display well-defined splicing patterns *in vivo*. In total, 571 binding sites met these criteria and were used in this analysis.

We investigated to which extent our *in vitro* landscape can be used to explain the *in vivo* binding of U2AF65 by performing *in silico* adjustment on the *in vitro* iCLIP landscape. This adjustment was applied based on the modeling of binding site affinities (K_d values) in different combinations of U2AF65 and RNA concentrations to produce a set of hypothetical landscapes (**Figure 13A**). The best fit was chosen as the hypothetical landscape showing the highest similarity to the *in vivo* iCLIP landscape. Despite overlaps between the best fit and *in vivo* binding landscapes, some discrepancies can be observed in many sites (**Figure 13B**). To quantify these discrepancies, we calculated the distances between the best fit and *in vivo* landscapes for each binding site as z-scores (**Figure 13A**). **Figure 13C** shows the distribution of the z-scores in different transcript regions, representing the regulatory events that occur *in*

vivo. Positive z-scores indicate binding sites that are stabilized, whereas negative z-scores represent binding sites that are cleared *in vivo*.

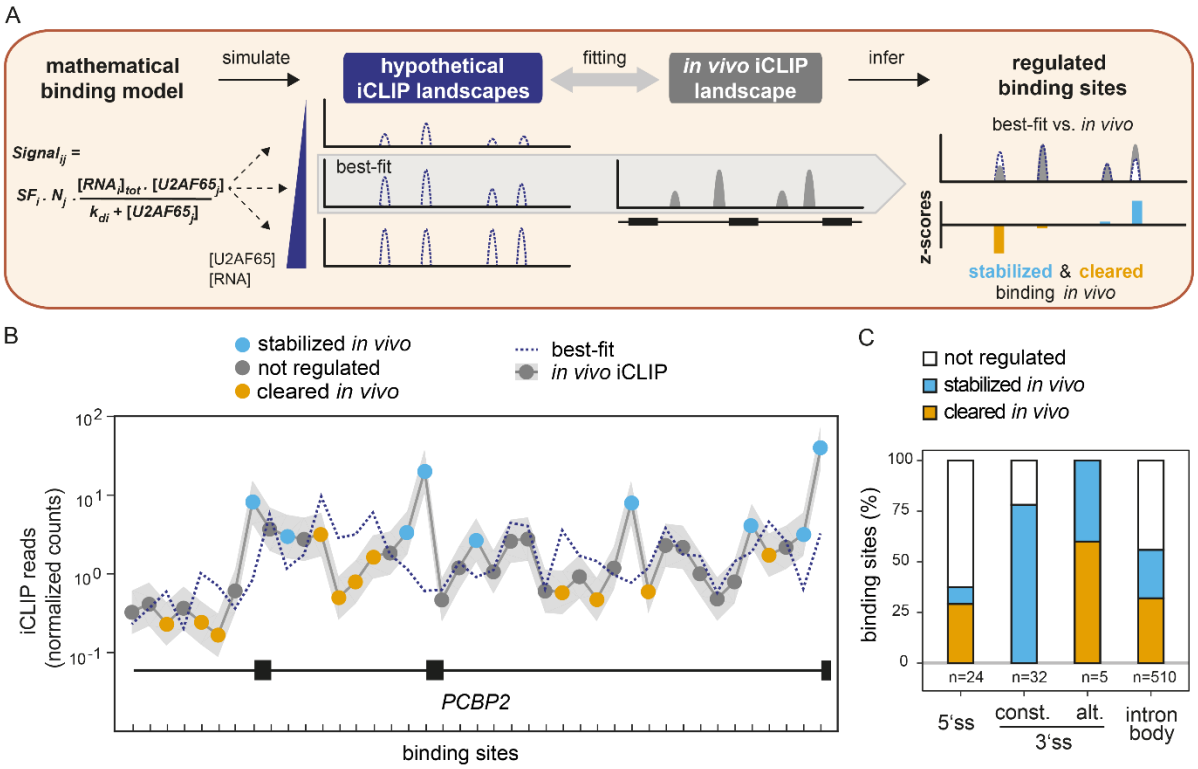


Figure 13. ***In vivo – in vitro* iCLIP comparative modeling**. **(A)** Schematic of *in vitro – in vivo* fitting to identify regulatory hotspots. **(B)** Comparison of best-fit (dotted blue line) and *in vivo* landscape (gray line) on *PCBP2* showing stabilized (blue) and cleared (orange) U2AF65 binding sites. Grey shadow represents standard deviation of *in vivo* iCLIP read counts from three independent replicates. **(C)** Plot showing the proportion of non-regulated (|z-score| < 1; white), stabilized (z-score > 1; blue) and cleared (z-score < -1; orange) binding sites in different transcript regions (5' splice site, constitutive or alternative 3' splice site, and intron body).

In our *in vitro* transcript region, we found that approximately 57% of U2AF65 binding sites are regulated (**Figure 13C**), underlining the major role of these regulations on shaping the U2AF65 binding landscape *in vivo*. At alternative 3′ splice sites, binding sites are both cleared and stabilized in similar proportions, supporting the flexibility of the isoform switch function during alternative splicing. Intriguingly, the binding sites regulated in constitutive 3′ splice sites were all stabilized *in vivo*. By contrast, binding sites at intronic and 5′ splice sites were heavily cleared *in vivo* (**Figure 13C**). The clearance of these binding sites supported the

idea of a proofreading mechanism that reduces the binding of U2AF65 to non–3′ splice site regions and ensures the enrichment at 3′ splice sites to support the role in splicing.

U2AF65 forms a heterodimer with U2AF35 and binds to SF1 to form the complex that defines the 3′ splice site in the early step of splicing (Selenko et al., 2003; Wu & Maniatis, 1993; Wu et al., 1999). Therefore, we were curious whether the absence of both co-factors contributed to weaker *in vitro* U2AF65 binding at 3′ splice sites. U2AF35 and SF1 bind to an AG dinucleotide and a branch point sequence, respectively. Thus, we checked whether the presence of a nearby AG or branch point is correlated with the stabilization of U2AF65 binding at 3′ splice sites. **Figure 14A & B** show that the presence of AG and the distance to branch point only mildly correlate with stabilization of the nearby U2AF65 binding sites; however, the effect is significantly lower than the stabilization at 3′ splice sites. This finding indicates that additional co-factor(s) may be involved in stabilizing U2AF65 binding at 3′ splice sites.

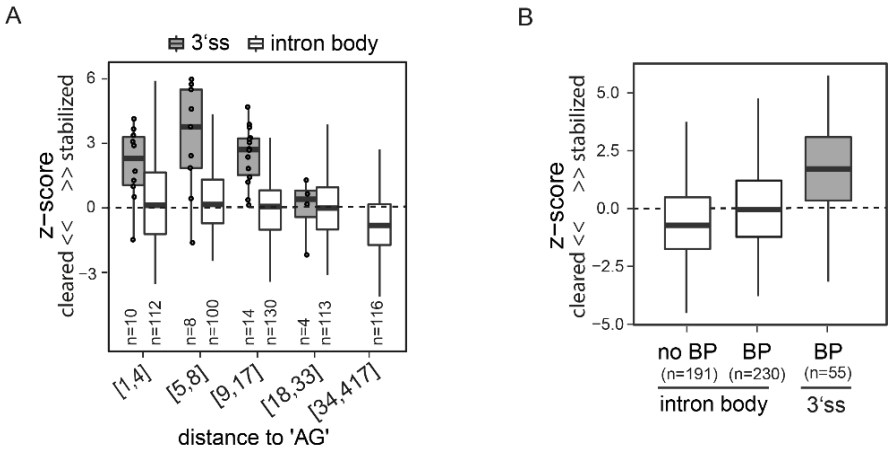


Figure 14. U2AF65 interactions with U2AF35 and SF1 only partially explain 3′ splice site stabilization *in vivo*. (A) Box plot showing the z-score distribution of U2AF65 binding sites at 3′ splice sites (gray) and in intron bodies (white). Binding sites were separated into six roughly equal-sized bins with increasing distance to the next AG dinucleotide (between 1 nt and 417 nt; indicated as ranges below). **(B)** Box plot as in (A) for binding sites without branch point (BP), with upstream BP motif, and with upstream BP motif and adjacent 3' splice site.

To cross-validate the strong stabilization of U2AF65 binding at 3′ splice sites, we performed a partial knockdown of *U2AF65* in HeLa cells (**Figure 15A**). The changes in cellular U2AF65 pool concentration was expected to decrease overall U2AF65 binding *in vivo*, especially on the weaker binding sites. Our knockdown data confirmed that U2AF65 binding sites at 3′ splice sites are the least affected by the knockdown among those transcript regions tested (**Figure 15B**), supporting the notion of strong stabilization at 3′ splice sites observed in the *in vitro*-*in vivo* comparison data.

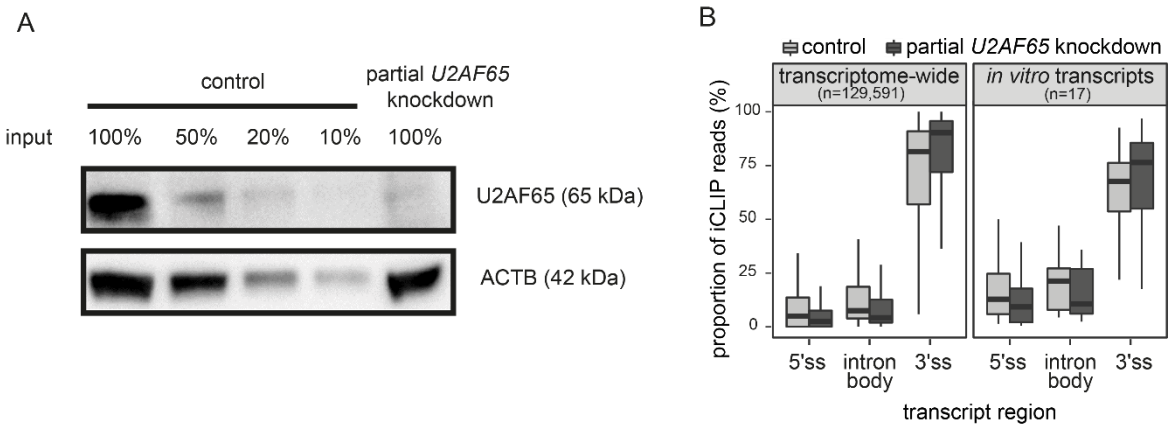


Figure 15. *In vivo* modulation of U2AF65 binding upon partial U2AF65 knockdown. (A) Western blot illustrating the U2AF65 protein level upon partial *U2AF65* knockdown. Actin beta (*ACTB*) was used as loading control. **(B)** Bar plot showing the proportion of *U2AF65 in vivo* iCLIP reads in control cells (light gray) and upon partial *U2AF65* knockdown (dark gray) for binding sites in different transcript regions. Analyses across transcriptome (left) as well as restricted to nine tested *in vitro* transcripts (right) are shown.

4.6 Machine learning identifies RBPs as potential U2AF65 regulators

Altogether, our data suggested that intense regulatory events control U2AF65 binding *in vivo* and seem to require additional co-factors. Therefore, we further sought to determine which of these co-factors that can potentially regulate U2AF65 binding. To this end, we used U2AF65 binding sites that were differentially regulated *in vivo* (|z-scores| > 1) based on the *in vitro–in vivo* U2AF65 binding comparison as an input using a machine learning approach (**Figure 16A**). We trained Random Forests to relate the presence of 4,224 sequence features (i.e., position-specific scoring matrices representing different RBP motifs, all possible 6-mers motifs, and positional information such as the relative location within the transcript) to the

direction of each binding site regulation and, thereby, correctly classify them into stabilized (z-scores > 1) or cleared (z-scores < -1) *in vivo* groups.

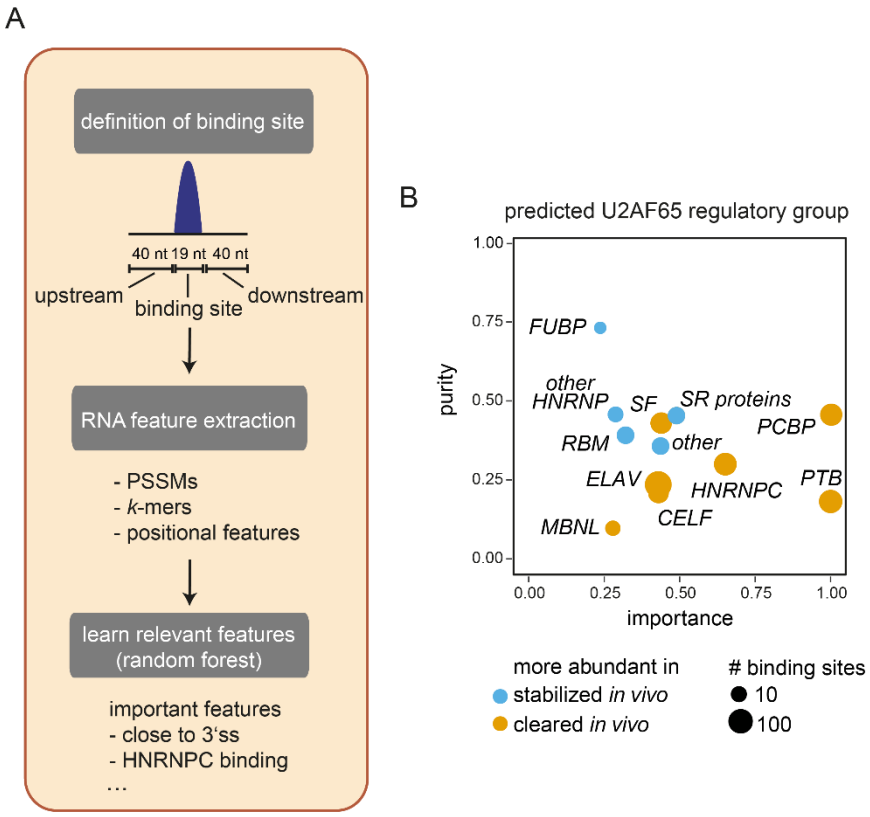


Figure 16. Identification of U2AF65 regulators with Random Forests. (A) Schematic workflow of the Random Forests approach that learns the most relevant features to classify U2AF65 binding sites into stabilized (z-score > 1, 151 sites) or cleared (z-score < -1, 173 sites) *in vivo*. **(B)** Twelve regulatory groups are identified as top candidates for *in vivo* U2AF65 regulation. Plot contrasting the relative importance and purity of collapsed regulatory groups from the top 100 features obtained by Random Forests analysis. Purity indicates specificity of association with a certain direction of regulation. Circle diameter represents scaled number of sites with predicted binding sites of a representative RBP from the group for the predominant direction of regulation (blue = stabilized *in vivo*, orange = cleared *in vivo*).

The machine learning approach produces a list of the features ranked by their importance for binding site classification. We selected the top 100 features based on the random forests result to perform further analysis. Among the top 100 features, we then

extracted the list of RBPs as potential regulators by focusing on either selecting features of a position-specific scoring matrix that directly appeared on the list or mapping the 6-mers features for a possible match of RBP motifs. We collapsed RBPs that had similar motifs into the same protein group (such as the paralogues PCPB1/2/3). This analysis yielded 12 protein groups considered to be potential regulators for U2AF65 (**Figure 16B**). The 12 protein groups consisted of regulators that we predicted to clear (such as hnRNPC, PTB, and PCBP) or stabilize (such as SR protein, FUBP, and RBM) U2AF65 binding *in vivo*.

4.7 Validation of predicted U2AF65 regulatory events *in vitro*

To first validate our prediction, we chose 11 different recombinant RBPs representing the different protein groups, which consist of both known (including hnRNPC, ELAVL1, PTBP1, and MBNL1) (König et al., 2010; Warf et al., 2009; Saulière et al., 2006; Izquierdo, 2008) and novel U2AF65 regulators (including FUBP1, RBM24, CELF6, KHDRBS1, PCBP1, SNRPA, and RBM41). We independently added recombinant purified RBPs to the *in vitro* iCLIP mix containing $U2AF65^{RRM12}$ and *in vitro* transcript set (nine transcripts) in the binding buffer. For the controls, we used *in vitro* iCLIP mix with the addition of BSA without any added co-factors. Independent $U2AF65^{RRM12}$ *in vitro* iCLIP libraries were then produced from the addition of each co-factor RBP to perform comparative analysis and detect *in vitro* $U2AF65^{RRM12}$ binding changes upon the co-factor additions.

The changes on $U2AF65^{RRM12}$ binding were calculated as $\log_2$ fold changes ($\log_2$FC), as represented in **Figure 17A**. Many U2AF65 binding sites are modulated upon the addition of the co-factor RBPs, and the majority of these modulations have a distinct pattern compared to the addition of BSA, indicating specific effects on $U2AF65^{RRM12}$ binding modulations by co-factor RBPs. In addition, some binding sites are modulated by multiple co-factor RBPs, suggesting that cross-interaction between co-factor RBPs may be important for producing the final functional regulatory events. As expected, suppression of U2AF65 binding can be observed upon the addition of several co-factors. This suppression is most likely the result of competition for the same binding site between the co-factors and U2AF65. Intriguingly, stabilization of U2AF65 binding was also observed upon the addition of several co-factors,

despite the lack of an RS domain in our recombinant U2AF65^{RRM12} known to mediate protein-protein interaction. This result suggested that stabilization of U2AF65 binding may be facilitated via the interaction of the co-factors with different domains of U2AF65. It is also possible that the stabilization by co-factors occurs indirectly through the formation of RNA secondary structures, which facilitates U2AF65 binding at some sites.

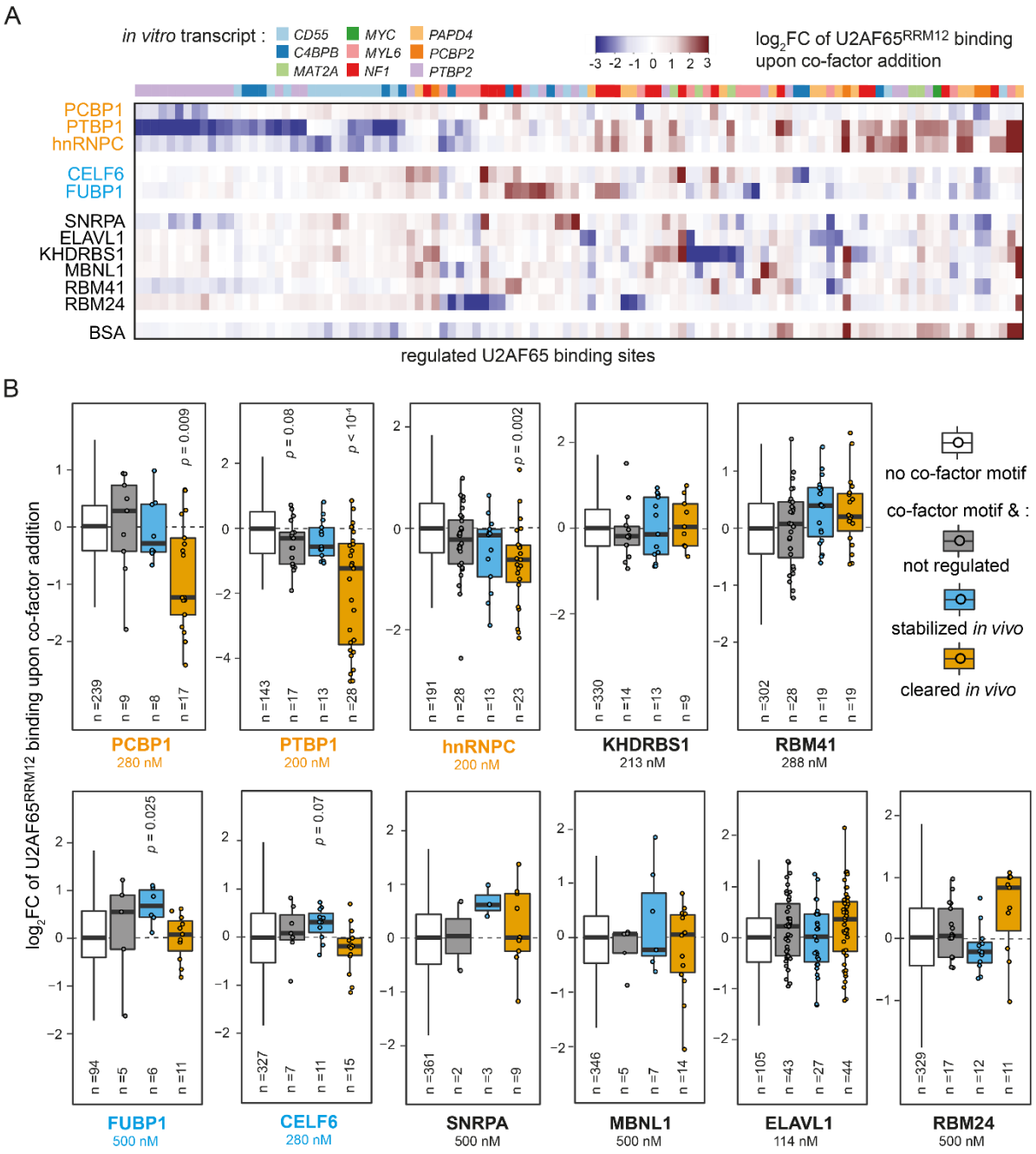


Figure 17. Co-factors change U2AF65RRM12 binding *in vitro*. (A) Heatmap showing log_2FC of normalized $U2AF65^{RRM12}$ read counts upon addition of co-factors ($U2AF65^{RRM12}$+co-factor/$U2AF65^{RRM12}$). *in vitro* transcripts indicated above **(B)** Bar plot showing log_2FC upon co-factor addition as in (A) for sites with no co-factor motif (white) as well as sites with co-factor motif and not regulated (gray, -0.5 < z-score < 0.5), stabilized *in vivo* (blue, z-score > 1), or cleared *in vivo* (orange, z-score < -1). Data were scaled such that log_2FC of sites without co-factor motif are centered around zero. Added co-factor concentrations are indicated in each panel. Adjusted *p*-values are given for all groups with false discovery rate (FDR) < 10% (two-sided Student's t-test, Benjamini-Hochberg correction, compared to binding sites without co-factor motif).

To relate the *in vitro* $U2AF65^{RRM12}$ binding modulation with our prediction, we grouped U2AF65 binding sites based on their model-predicted regulatory events, as shown in **Figure 17B**. We found that hnRNPC, PTBP1, and PCBP1 decreases $U2AF65^{RRM12}$ binding at the binding sites harboring the co-factor motifs and predicted to be cleared *in vivo*. These findings indicate that these RBPs are U2AF65 binding suppressors, supporting previous studies on hnRNPC and PTBP1 (König et al., 2010; Saulière et al., 2006). By contrast, CELF6 and FUBP1 significantly increased U2AF65 binding at the predicted binding sites, revealing a potentially novel role of these RBPs as enhancers of U2AF65 binding. For the other co-factor RBPs, we could not find any significant enrichment on the direction of the regulatory effects. This lack of enrichment may be due to the small number of binding sites predicted to be regulated in our *in vitro* transcript set (such as for MBNL1 and SNRPA) or the lack of additional factor(s) or domain(s) that may be necessary for regulation by these co-factor RBPs.

4.8 *In vivo* regulation by hnRNPC can be recapitulated *in vitro*

As a well-established U2AF65 regulator, hnRNPC suppresses U2AF65 binding via competition for uracil-rich regions of binding sites (Zarnack et al., 2013). To further explore the relevance of our validated regulatory events to *in vivo* regulation, we used published iCLIP data for a comparison. In addition, we performed additional *in vitro* iCLIP experiments by adding different concentrations of recombinant hnRNPC1 to the *in vitro* iCLIP mixes to evaluate the effect of hnRNPC concentrations on U2AF65 binding *in vitro*.

Figure 18A shows the $\log_2$FC of U2AF65^{RRM12} binding upon the addition of different concentrations of recombinant hnRNPC1. As we observed previously, hnRNPC has a clear enrichment toward clearance of U2AF65 binding at the sites predicted to be regulated by hnRNPC based on the motif prediction. This effect was observed in all three concentrations of added hnRNPC1 but in different strengths, indicating that the competition between hnRNPC and U2AF65 is concentration dependent. When we used U2AF65-regulated sites that are upregulated upon *HNRNPC* knockdown from published *in vivo* U2AF65 iCLIP data, we observed comparable clearance effects to our *in vitro* assay (**Figure 18A,** Zarnack et al., 2013). In addition, we repeated the same analysis by using *in vivo* hnRNPC iCLIP data to define hnRNPC binding sites instead of using motif prediction. Here, the same pattern of U2AF65 binding clearance was observed in the binding sites bound by hnRNPC *in vivo* (**Figure 18C**). One example can be illustrated by the iCLIP data in *CD55* alternative exon 10, which is known to be regulated via hnRNPC-U2AF65 competition, as shown in **Figure 18B**. To finally compare our *in vitro–in vivo* regulation by hnRNPC, we plotted changes of U2AF65^{RRM12} binding *in vitro* upon the addition of hnRNPC1 against the changes *in vivo* upon *HNRNPC* knockdown. Here, we showed that both datasets demonstrated a significant correlation of hnRNPC regulatory effect with U2AF65 binding. In conclusion, our data suggested that we could recapitulate the competition between hnRNPC and U2AF65 *in vivo* by using our *in vitro* co-factor assays.

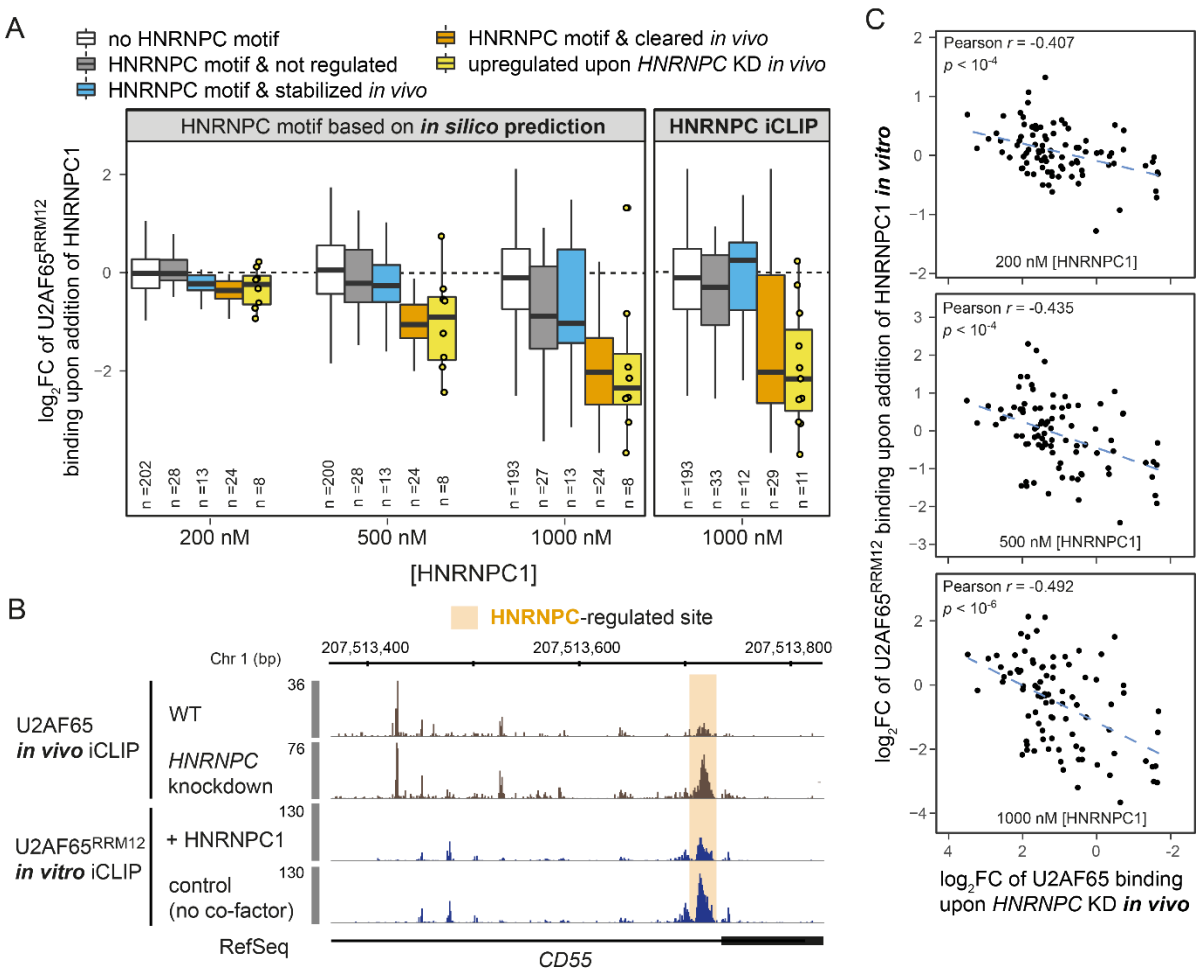


Figure 18. Recapitulation of regulation by hnRNPC *in vitro*. (A) hnRNPC1 suppresses U2AF65^{RRM12} binding specifically at sites that are downregulated *in vivo* and show overlapping hnRNPC binding according to *in silico* predictions (left panel) or hnRNPC *in vivo* iCLIP data (right panel). Boxplots of log_2-transformed fold changes (log_2FC) of normalized U2AF65^{RRM12} read counts in *in vitro* iCLIP co-factor assays with different concentrations of recombinant hnRNPC1 (200, 500 and 1000 nM) over a control with U2AF65^{RRM12} alone. U2AF65 binding sites are subdivided into sites without hnRNPC (white) and sites with hnRNPC but not regulated *in vivo* (z-score < |0.5|; gray), stabilized *in vivo* (z-score > 1; blue) or cleared *in vivo* (z-score < -1; orange). Additional box summarizes all sites with decreased U2AF65 binding upon *HNRNPC* knockdown (KD) *in vivo* (log_2FC < -1; yellow) (Zarnack et al. 2013). **(B)** Genome browser view of U2AF65 *in vivo* iCLIP data (brown) from control and *HNRNPC* knockdown HeLa cells (Zarnack et al. 2013), as well as *in vitro* iCLIP data (blue) for U2AF65^{RRM12} alone (control) and upon addition of recombinant hnRNPC1 at hnRNPC-regulated alternative exon of *CD55*. Orange shading highlights hnRNPC-regulated site. **(C)** Scatter plots comparing log_2FC upon *in vitro* addition of hnRNPC1 (hnRNPC1+U2AF65^{RRM12}/U2AF65^{RRM12}) with log_2FC of U2AF65 *in vivo* iCLIP read counts upon *HNRNPC* KD (KD/control) for three different concentrations of recombinant hnRNPC1. Pearson *r* and associated *p*-values are given.

4.9 *PTBP2* exon 10 alternative splicing regulation by PTBP1 and FUBP1

The regulation of *PTBP2* exon 10 is a well-studied alternative splicing event that is important to post-transcriptionally controlling the expression level of the *PTBP2* gene in the cell. The homolog protein of this gene product, PTBP1, is one of the known regulators of *PTBP2* exon 10 alternative splicing (Boutz et al., 2007; Xue et al., 2009). This regulation occurs via competition with U2AF65 at the 3′ splice site of the alternative exon, where the skipping of this exon will lead to the production of an isoform that is directed for non-sense mediated RNA decay.

Table 3. Mutations list for minigenes construct of *PTBP2* exon 10 splicing assay.

Mutant	Mutated region	Mutation
$BS1_{del}$	chr1:97271911-97271945	deleted
$BS2_{del}$	chr1:97271587-97271751	deleted
$BS2_{sub}$	chr1:97271588-97271751	CTATATTTTATTTTGTTTTTGTTCCCCAATTCCTTATTTTTTTCTTCTGCATTGCTGTTTCCCTTCCCCATTTCATCCTTTTCCCTGTGTGTTCACCTTCCCTTTCCTTGTCTTTTCCCAAATGCCCATTCCCTTCCCTGTCTTATCCTTTATTTTCCTTGTCC → TAGTTAACCTTTGCAGCATTTGTTTACAGTTTACAGTTCC (chr3:186506484-186506523)
ΔBS3	chr1:97271420-97271432	ATGCTTTCCTTCC → AAGCTATCGTTAC

PTBP2 exon 10 was part of our *in vitro* transcript set. Thus, we checked whether we could observe PTBP1 and U2AF65 competition in our *in vitro* co-factor assays. We found that the addition of recombinant PTBP1 decreases U2AF65 binding at the binding sites upstream from the 3′ splice site of *PTBP2* exon 10 (shown as BS1 in **Figure 19**). In addition, *in vivo* iCLIP data (unpublished data) showed that PTBP1 binds to BS1 and seems to suppress U2AF65 binding at this site. This observation supported the regulatory event captured by our *in vitro* co-factor assays. To validate this finding, we designed a *PTBP1* minigene reporter, which carries a mutation in BS1 that eliminates the PTBP1 binding site but still preserves the

nearby U2AF65 binding site to allow for recognition of the corresponding 3′ splice site (**Table 3**, **Figure 19**). This mutation releases suppression by PTBP1 and increases the inclusion of *PTBP2* exon 10 to 100% (**Figure 20A & B**). The same effect was observed when we depleted the PTBP1 level by using siRNA, indicating that BS1 indeed regulator of PTBP1 important for controlling alternative splicing of *PTBP2* exon 10 via competition with U2AF65.

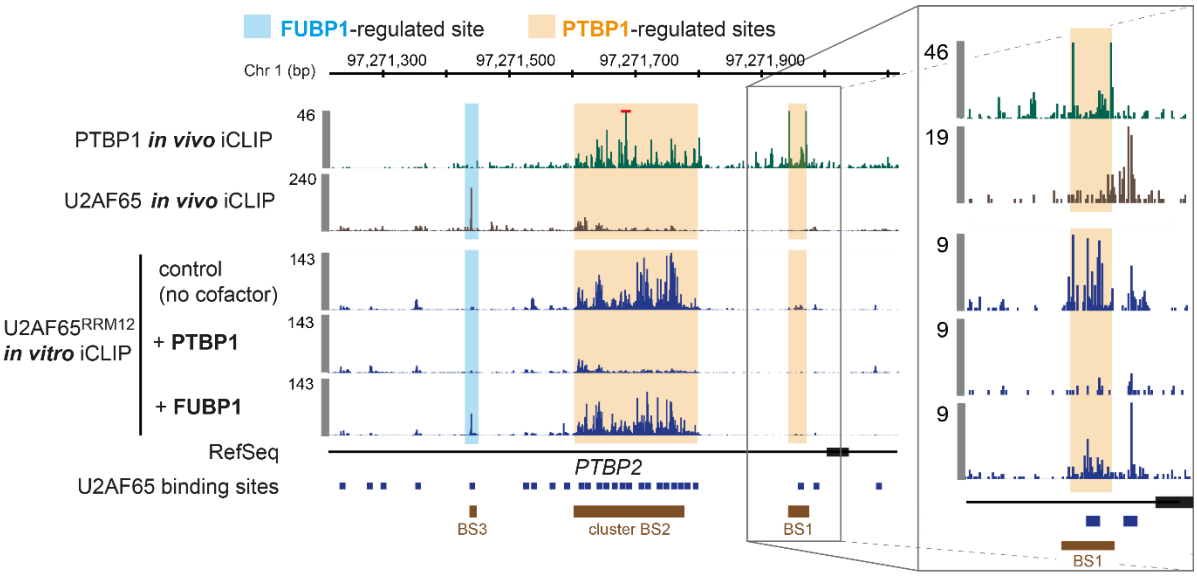


Figure 19. *In vitro* experiment captures in vivo regulation at PTBP2 exon 10. Genome browser view of *in vivo* iCLIP for PTBP1 (green) and U2AF65 (brown), as well as *in vitro* iCLIP for U2AF65^{RRM12} alone and upon addition of recombinant FUBP1 and PTBP1 proteins upstream of *PTBP2* exon 10. Regulated U2AF65 sites are marked as BS1-3.

We observed another cluster of U2AF65 binding sites upstream from BS1 (cluster BS2) that were significantly cleared upon the addition of PTBP1 *in vitro* and bound by PTBP1 *in vivo* (**Table 3**, **Figure 19**). We wondered whether U2AF65 binding in this region had a functional relevance in the alternative splicing regulation of *PTBP2* exon 10. To check this, we transfected HeLa cells with the *PTBP2* minigene that contained a substitution of cluster BS2 with U2AF65 binding sites not regulated by PTBP1 to allow for U2AF65 binding in this position. Intriguingly, this substitution caused a decreased in the level of *PTBP2* exon 10

inclusion. Although the changes in the inclusion level was only slightly affected, the opposite effect was observed when we completely deleted cluster BS2, thus removing U2AF65 binding. This finding indicated that PTBP1 regulation at cluster BS2 contributes to certain extent to the regulation of *PTBP2* exon 10 alternative splicing. Further increases in exon inclusion were observed when we depleted PTBP1 (**Figure 20A & B**), suggesting that cluster BS2 is more auxiliary than the main PTBP1 regulatory hotspot that controls alternative splicing in this region.

In contrast to PTBP1, FUBP1 enhanced U2AF65 binding at the site downstream from cluster BS2 *in vitro* (BS3, **Table 3**, **Figure 19**). Intriguingly, U2AF65 binding in this site was stabilized *in vivo*, compared to that in the *in vitro* binding landscape. Moreover, the knockdown of *FUBP1* decreased the inclusion of *PTBP2* exon 10, supporting the notion of FUBP1 regulation in this region observed in our *in vitro* co-factor assays. To investigate the role of BS3 in the regulation of *PTBP2* exon 10 alternative splicing, we mutated BS3 to remove the U2AF65 binding motif. We transfected the minigene containing this mutation into HeLa cells and monitored the splicing pattern of *PTBP2* exon 10. Similar to cluster BS2, we found that removing U2AF65 binding at this binding site slightly increases exon 10 inclusion (**Figure 20A & B**). However, the knockdown of *FUBP1* still significantly affected the splicing of *PTBP2* exon 10, suggesting that FUBP1 may regulate *PTBP2* exon 10 alternative splicing via additional sites.

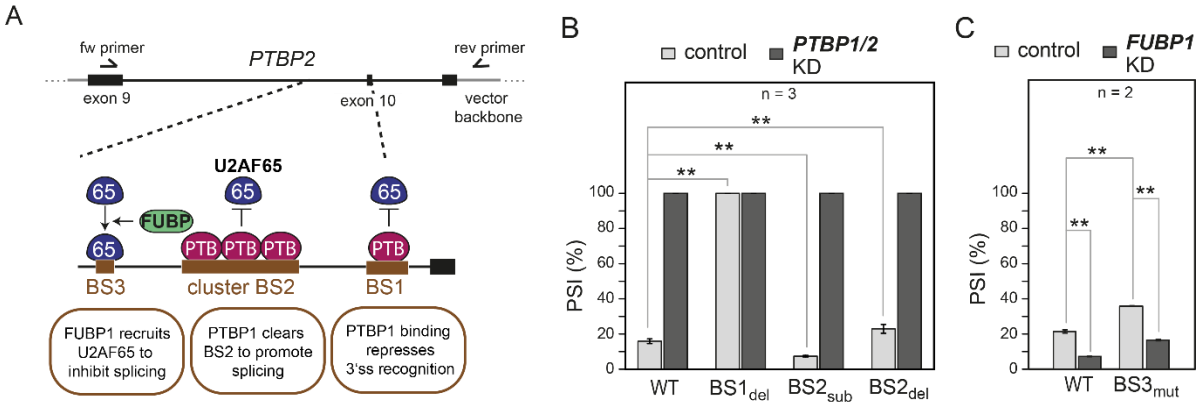


Figure 20. Minigene reporter assays confirm PTBP2 exon 10 regulation by PTBP1 and FUBP1 *in vivo*. (A) Top**:** RT-PCR primers to measure splicing changes. Bottom: Schematic model of U2AF65 regulation at BS1-3 and its impact on inclusion of *PTBP2* exon 10. **(B,C)** Inclusion of *PTBP2* exon 10 is conjointly regulated by BS1, BS2 and BS3. Bar plots showing *PTBP2* exon 10 inclusion (depicted as 'percent spliced in', PSI) in control (light gray) and *PTBP1/2* **(B)** or *FUBP1* **(C)** knockdown (dark gray) HeLa cells. Minigene constructs include wild-type (WT) and four mutated versions with BS1 deletion ($BS1_{del}$), BS2 deletion ($BS2_{del}$), substitution of BS2 with a U2AF65 binding site that is not regulated by PTBP1 ($BS2_{sub}$), and BS3 mutation that eliminates the U2AF65 recognition motif ($BS3_{mut}$). * *p*-value < 0.05, ** *p*-value < 0.01 (two-sided Student's t-test). Error bars represent standard deviation of the mean.

Altogether, here, we showed that our *in vitro* iCLIP data can help to disentangle complex regulatory mechanisms of *PTBP2* exon 10 and, therefore, provide us with an additional layer of information about the regulation of alternative splicing in this gene.

4.10 Relevance of predicted regulation to splicing decision *in vivo*

We ultimately sought to determine to what extent our regulatory prediction can be used to anticipate the outcomes of alternative splicing *in vivo*. To this end, we performed the knockdown of all 11 RBPs used in the *in vitro* co-factor assays and measured their effects on the endogenous splicing outcomes of four alternative exons present in our *in vitro* transcript set (*PTBP2* exon 10, *MYL6* exon 6, *CD55* exon 10, and *PCBP2* exon 12). We correlated the alternative splicing outcomes in each knockdown with the *in vitro* U2AF65 binding modulations captured in the *in vitro* co-factor assays within 600 nucleotides upstream from the 3′ splice site of each alternative exon (**Figure 21**).

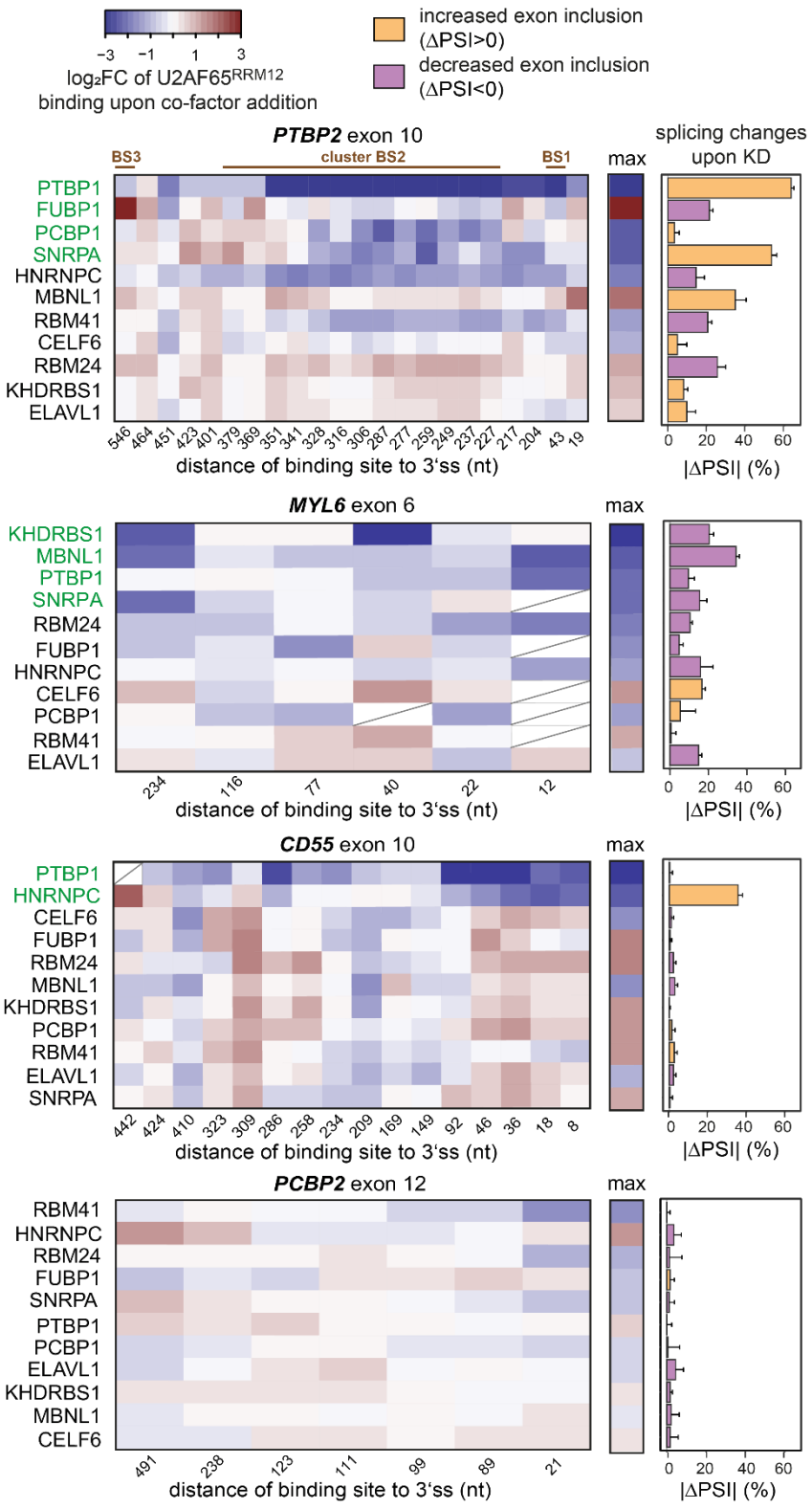

−3 −1 1 3
log₂FC of U2AF65RRM12
binding upon co-factor addition
increased exon inclusion
(ΔPSI>0)
decreased exon inclusion
(ΔPSI<0)
PTBP2 exon 10
BS3
cluster BS2
BS1
max
splicing changes
upon KD
PTBP1
FUBP1
PCBP1
SNRPA
HNRNPC
MBNL1
RBM41
CELF6
RBM24
KHDRBS1
ELAVL1
546 464 451 423 401 379 369 351 341 328 316 306 287 277 259 249 237 227 217 204 43 19
distance of binding site to 3'ss (nt)
0 20 40 60
|ΔPSI| (%)
MYL6 exon 6
max
KHDRBS1
MBNL1
PTBP1
SNRPA
RBM24
FUBP1
HNRNPC
CELF6
PCBP1
RBM41
ELAVL1
234 116 77 40 22 12
distance of binding site to 3'ss (nt)
0 20 40 60
|ΔPSI| (%)
CD55 exon 10
max
PTBP1
HNRNPC
CELF6
FUBP1
RBM24
MBNL1
KHDRBS1
PCBP1
RBM41
ELAVL1
SNRPA
442 424 410 323 309 286 258 234 209 169 149 92 46 36 18 8
distance of binding site to 3'ss (nt)
0 20 40 60
|ΔPSI| (%)
PCBP2 exon 12
max
RBM41
HNRNPC
RBM24
FUBP1
SNRPA
PTBP1
PCBP1
ELAVL1
KHDRBS1
MBNL1
CELF6
491 238 123 111 99 89 21
distance of binding site to 3'ss (nt)
0 20 40 60
|ΔPSI| (%)

Figure 21. *In vitro* changes predict *in vivo* regulation of alternative splicing. Left panel: Heatmaps showing $\log_2$FC of normalized *in vitro* U2AF65^{RRM12} read counts upon co-factor addition (U2AF65 ^{RRM12}+co-factor/U2AF65^{RRM12}) for binding sites within 600 nt of the 3' splice sites of four alternative exons (*PTBP2* exon 10, *MYL6* exon 6, *CD55* exon 10 and *PCBP2* exon 12; note that only 300 nt are shown for *MYL6* exon 6 corresponding to the maximal length of the preceding intron). Co-factors in each heatmap are ordered by the maximum change in U2AF65^{RRM12} binding at any site *in vitro* (summarized in the mid panel). Binding sites that were excluded due to low coverage are marked in gray. Right panel: Bar chart showing absolute changes in exon inclusion ('Percent Spliced In', PSI) *in vivo* upon KD of individual co-factors. Colors indicate direction of splicing change (orange, upregulation upon KD; purple, downregulation). Based on their maximum effect on U2AF65^{RRM12} binding *in vitro*, co-factors were subdivided into "*in vitro* U2AF65 regulators" (green, maximal $|\log_2 FC| > 2$ in U2AF65^{RRM12} binding upon co-factor addition) and "non-regulators" (white, maximal $|\log_2 FC| \leq 2$). Error bars represent standard deviation of the mean (n=3).

We found that the strength of the regulatory events measured in our *in vitro* co-factor assays were well correlated with the changes in alternative splicing decisions (**Figure 21**). This correlation was observed especially in the alternative exons that are strongly regulated *in vitro*, such as *PTBP2* exon 10, and *MYL6* exon 6. Moreover, the absence of *in vitro* regulation upstream from *PCBP2* exon 12 was well translated into the lack of splicing changes on this exon *in vivo* upon knockdown of its co-factors. To quantify this correlation, we grouped the co-factor RBPs into non-U2AF65 and U2AF65 regulators based on their strengths of U2AF65 binding modulations *in vitro* (**Figure 22**). Here, we showed that in all alternative exons that were tested, the knockdown of U2AF65 regulators resulted in stronger changes on the alternative splicing decision than the non-U2AF65 regulators. However, the direction of the alternative splicing changes cannot be inferred from the *in vitro* co-factor assays. This inference cannot be made because U2AF65 binding regulation per se is not predictive of the direction of the alternative splicing decision. Instead, the effect of U2AF65 binding modulations on alternative splicing decision is very dependent on the location of the regulatory events, as we showed in previous experiments with *PTBP2* exon 10 (**Figure 20A & B**).

In conclusion, we showed that the intensity of the *in vitro* U2AF65 regulatory events upstream from the alternative exons correlates with changes in the inclusion level of the corresponding alternative exons *in vivo* upon the knockdown of different co-factors, indicating

that, *in vitro* regulatory scenarios by co-factors can represent cellular regulation in the *in vivo* context.

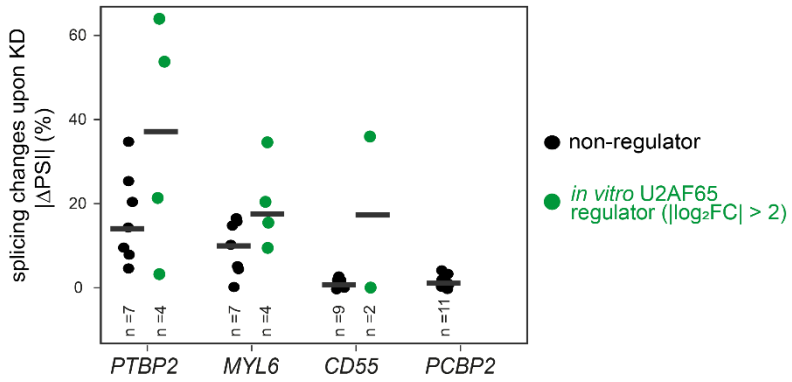


Figure 22. Predicted regulator RBPs affects *in vivo* alternative splicing decision. Bar plot showing ΔPSI upon co-factor KDs for four alternative exons.

5. DISCUSSION

5.1 *In vitro* iCLIP measures binding site affinity in a natural RNA context

Measuring binding site affinity has been a routine approach to understand molecular codes that drive the behavior of RNA-protein interactions in the cell. *In vitro* techniques are superior for this purpose due to the nature of the protocol with a defined system, which allows for better accuracy and precision of the affinity measurement. Several high-throughput *in vitro* techniques, such as RNA Bind-n-Seq and HTS-EQ (Lambert et al., 2014; Jain et al., 2017), have been routinely used to measure affinity of RNA-protein interactions. However, they employ short RNA sequences in their protocol, which may overlook the contribution of surrounding sequences in a longer RNA context in the *in vivo* system.

Here, we developed an *in vitro* iCLIP protocol to measure the binding affinities of multiple U2AF65 binding sites simultaneously, which increases the throughput of such measurement. One critical aspect that we introduce in the *in vitro* iCLIP is the use of long transcripts representing parts of human transcriptome (~2K nt on average). This aspect allows us to evaluate the influence of the surrounding sequences relevant to the *in vivo* context, such as the formation of secondary structures shown to be an important factor for modulating RNA-protein interactions both *in vivo* and *in vitro* (Lambert et al., 2014; Luo et al., 2016).

In this study, we used U2AF65, an RBP that prefers binding to the Py tract in a single-stranded RNA molecule. To capture most possible U2AF65 binding scenarios to single-stranded RNA, we reduced the RNA secondary structures by pre-heating the transcripts used in the *in vitro* iCLIP. Nevertheless, we were still able to show that incorporating the secondary structure information improves the reliability of the measurement of U2AF65 binding site affinities, indicating that some secondary structures may be reformed after the heat treatment due to the absence of denaturing agents important for preventing the refolding of RNA (Lambert & Draper, 2012; Lehrach et al., 1977). It is also possible that some secondary structures on the RNA are more thermostable and serve a specific regulatory purpose (Wan et al., 2012). The regulation of U2AF65 binding through RNA secondary structures has been shown to exist in the regulation of cardiac troponin T (*TNNT2*) alternative splicing (Warf et

al., 2009). Therefore, our data suggested that such regulation through RNA secondary structure is not exclusive but rather a widespread phenomenon for U2AF65 binding.

5.2 Calibration of the *in vivo* iCLIP signal

iCLIP and other crosslinking-based techniques represent the distribution of protein occupancies along the transcript as crosslinking events that are captured by the protocol. Inevitably, variability in crosslinking efficiencies of different binding sites that occur due to different sequence compositions often introduce technical biases in the iCLIP data (Sugimoto et al., 2012). Therefore, when quantitative analysis of the iCLIP data is preferred, the sensitivity of the method can be improved by taking these biases into account. Based on *in vitro* titration experiments, our K_d modeling quantitatively estimated the crosslinking and other biases that are derived from sequence variance on the individual binding site as scaling factors (*SF*s, **Figure 10**). This estimation allows for the correction of the iCLIP signal and, therefore, provides a more precise downstream quantification. In principle, a list of the scaling factors and their corresponding binding site sequences could be used as an input to further derive general calibration factors to be implemented directly based on the specific sequence composition of the binding sites for a particular RBP. Similar approaches for incorporating technical biases in different techniques such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) have been shown to improve the interpretation of the binding data (Nettling et al., 2016; Yardımcı et al., 2014). Furthermore, such information would be valuable to understand the biophysical aspect of crosslinking-based techniques.

5.3 U2AF65 binding is stabilized at 3′ splice site and cleared in intronic region

From our comparative study, we found an extensive stabilization of U2AF65 binding at 3′ splice sites *in vivo*. Intriguingly, such regulation was not limited to a few subsets of these sites but rather the majority of the U2AF65 binding sites at 3′ splice sites, especially in front of constitutive exons (**Figure 13C**). These findings indicated that the recruitment of U2AF65 to 3′ splice sites by co-factors is a general rule rather than an exception in the recognition of

3′ splice sites by U2AF65. This rule provides another layer of control in 3′ splice site recognition that could explain the robustness of U2AF65 binding when U2AF65 concentration was depleted *in vivo*. In support to this notion of robust U2AF65 binding at 3′ splice sites, a similar observation was also reported by Shao et al. who found only ~6% of the detected alternative exons change their inclusion level upon U2AF65 knockdown (Shao et al., 2014). Considering the importance of U2AF65 binding at 3′ splice sites, such a mechanism may be necessary to ensure robust splicing under different cellular conditions that compromise the abundance of the U2AF65 pool in the cell.

Another profound regulation that we observed in our dataset was the clearance of U2AF65 binding in intronic regions (**Figure 13C**). This finding supports the notion of a proofreading mechanism to minimize U2AF65 binding in "unproductive" binding sites and direct it to 3′ splice sites. Both our experimental datasets (**Figure 20**) and previous studies have shown that U2AF65 binding in intronic regions could affect the splicing outcome of the downstream exon, most likely by competing with U2AF65 binding sites at 3′ splice sites (Shao et al., 2014). Therefore, clearance of intronic binding by co-factors may be as essential as the stabilization at 3′ splice sites for the functional recruitment of U2AF65. Such examples of proofreading for U2AF65 binding have been shown in hnRNPA1 and DEK by selectively suppressing U2AF65 binding at the Py tracts not followed by the 3′ splice site landmark, AG dinucleotide (Tavanez et al., 2012; Soares et al., 2006).

5.4 Identification of U2AF65 regulators

We screened and identified RBPs as potential U2AF65 binding regulators. Among them, we identified hnRNPC as a known suppressor of U2AF65 binding via direct competition for RNA sequences (Zarnack et al., 2013). Furthermore, we showed that *in vivo* competition between hnRNPC and U2AF65 could be recapitulated in our *in vitro* setup, supporting a previous finding that this competition requires no additional factors and occurs in a direct fashion (Zarnack et al., 2013). In addition, a second known U2AF65 binding suppressor that was identified in the screening was PTBP1 (Saulière et al., 2006). As previously described, we showed that PTBP1 strongly competes with U2AF65 at the 3′ splice

site of *PTBP2* exon 10 and, therefore, inhibits the exon inclusion during the alternative splicing decision.

From our screening, PCBP1 was identified as a third suppressor of U2AF65 binding. Although PCBP1 has been known as a splicing regulator, its role has been primarily limited to the regulation of *CD44* alternative splicing, which controls the production of the tumor-promoting isoform of this gene (Meng et al., 2007; Tripathi et al., 2016). In addition, it is unclear how PCBP1 regulates the alternative splicing event of *CD44*. Therefore, our screening results suggested a potential novel splicing regulatory mechanism of PCBP1 via suppression of U2AF65 binding.

Using *in vitro* co-factor assays, we also identified novel enhancers of U2AF65 binding such as CELF6 and FUBP1. Because recombinant U2AF65^{RRM12} lacks the RS domain implicated in protein-protein interactions (Boucher et al., 2001), such enhancement of U2AF65 binding must be accomplished via additional interaction interfaces. Future *in vitro* co-factor assays could involve truncated versions of U2AF65^{RRM12} to better understand the interaction domain(s) required for the regulation mediated by these co-factors. It is also possible that these enhancers promote U2AF65 binding by stabilizing certain RNA secondary structures. Such splicing enhancing scenarios that depend on the formation of RNA secondary structures have been observed for different splicing factors, such as B52, SRp55, and NOVA-1 (Buckanovich & Darnell, 1997; Shi et al., 1997; Nagel et al., 1998).

In splicing regulation, little is known about the role of CELF6 and FUBP1. CELF6 belongs to the CUG-BP and ETR-3-like factor (CELF) protein family and is only abundantly expressed in the kidney, brain, and testis (Ladd et al., 2004). Therefore, it might be of interest to further investigate the knockdown effect of this factor on alternative splicing in different cell lines to better understand the interplay with U2AF65. In addition, other family members had already been shown to interfere with U2AF65 binding at *neurofibromatosis type I* (*NF1*) exon 23a (Barron et al., 2010). However, CELF6 was the only CELF protein that did not change the inclusion of this exon upon knockdown, and its role thus remains enigmatic. Despite its involvement in splicing regulation that is only beginning to be understood (Jacob et al., 2014; Li et al., 2013; Miro et al., 2015), FUBP1 is a well-studied single-stranded DNA

binding protein that plays an important role as a transcription factor (Chung & Levens, 2005; Duncan et al., 1994; Rabenhorst et al., 2009). Thus, our data support adding FUBP1 to the list of proteins with dual roles in transcription and splicing and further suggest that these regulatory processes are intimately connected (Han et al., 2017). Furthermore, recent publications have described FUBP1 binding to double-stranded RNA structures that may partially unfold the structures *in vivo* via its helicase activity (Li et al., 2013; Kralovicova & Vorechovsky, 2017). Therefore, it is tempting to speculate that FUBP1 may open critical binding sites that are buried in secondary structure to provide access for U2AF65.

5.5 *In vitro* iCLIP disentangles complex regulatory mechanisms *in vivo*

Alternative splicing is regulated by the concerted action of hundreds of RBPs that form a complex functional network. One type of such interaction has been proposed to be achieved via overlapping binding sites of RBPs that allows for multiple ways to regulate a particular splicing event in different biological contexts, such as different cell types and cellular conditions (Goren et al., 2010). By comparing the results of *in vitro* co-factor assays to the knockdown in living cells, we demonstrated that *in vitro* U2AF65^{RRM12} binding modulations upon co-factor additions predict knockdown-induced alternative splicing changes. This evidence supports the notion that U2AF65 binding is a pioneering event that controls spliceosome recruitment and alternative splicing decisions in living cells. Nevertheless, some discrepancies were still observed on individual cases, such as PTBP1 regulation at the *CD55* alternative exon. PTBP1 induces strong changes in U2AF65 binding *in vitro* but does not change the inclusion of the corresponding downstream alternative exon *in vivo*. This difference can be explained by the overlapping regulatory sites between PTBP1 and hnRNPC, creating competition during alternative splicing decisions. In our *in vivo* setup, we showed that hnRNPC dominates the regulatory function of the alternative exon; however, it would be of interest to see if such a regulatory event can be modulated by changing the cellular level ratio of PTBP1 and hnRNPC. Although direct investigation of these overlapping regulatory events may be challenging *in vivo*, nevertheless *in vitro* reconstruction can be done in the future by performing *in vitro* iCLIP of U2AF65 with a different ratio of added hnRNPC and PTBP1.

In conclusion, using *in vitro* iCLIP, we showed that *in vivo* U2AF65 binding is strongly shaped by auxiliary RBPs that direct U2AF65 to 3′ splice sites to control alternative splicing (**Figure 23**). In the future, the power of this method can be improved by performing *in vitro* iCLIP on a transcriptome-wide scale using extracted cellular RNAs. Furthermore, this approach can be extended to other RBPs, thus providing a resource for RNA-protein interaction study and further insights into the complexity of mRNP assembly.

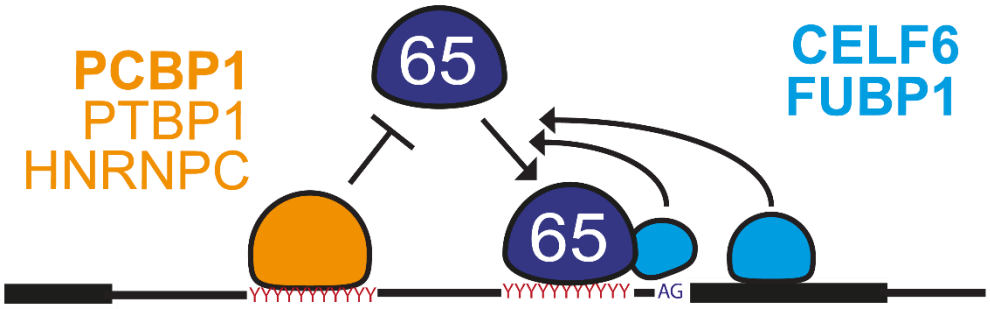


Figure 23. Working model of U2AF65 recruitment to 3′ splice site. U2AF65 is directed to 3′ splice site via co-factors-mediated stabilization at 3′ splice site and binding clearance at intronic region.

6. REFERENCES

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. Nature communications *7*, 10950.

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S., and Tuschl, T. (2012). Identification of RNA-protein interaction networks using PAR-CLIP. Wiley interdisciplinary reviews. RNA *3*, 159–177.

Banerjee, H., Rahn, A., Gawande, B., Guth, S., Valcárcel, J., and Singh, R. (2004). The conserved RNA recognition motif 3 of U2 snRNA auxiliary factor (U2AF65) is essential in vivo but dispensable for activity in vitro. RNA *10*, 240–253.

Barron, V.A., Zhu, H., Hinman, M.N., Ladd, A.N., and Lou, H. (2010). The neurofibromatosis type I pre-mRNA is a novel target of CELF protein-mediated splicing regulation. Nucleic acids research *38*, 253–264.

Berget, S.M. (1995). Exon recognition in vertebrate splicing. The Journal of biological chemistry *270*, 2411–2414.

Blanchette, M., and Chabot, B. (1999). Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. The EMBO journal *18*, 1939–1952.

Boucher, L., Ouzounis, C.A., Enright, A.J., and Blencowe, B.J. (2001). A genome-wide survey of RS domain proteins. RNA *7*, 1693–1701.

Bourgeois, C.F., Popielarz, M., Hildwein, G., and Stevenin, J. (1999). Identification of a bidirectional splicing enhancer: differential involvement of SR proteins in 5' or 3' splice site activation. Molecular and cellular biology *19*, 7347–7356.

Boutz, P.L., Stoilov, P., Li, Q., Lin, C.-H., Chawla, G., Ostrow, K., Shiue, L., Ares, M., and Black, D.L. (2007). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes & development *21*, 1636–1652.

Buckanovich, R.J., and Darnell, R.B. (1997). The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. Molecular and cellular biology *17*, 3194–3201.

Buenrostro, J.D., Araya, C.L., Chircus, L.M., Layton, C.J., Chang, H.Y., Snyder, M.P., and Greenleaf, W.J. (2014). Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nature biotechnology *32*, 562–568.

Campbell, Z.T., Bhimsaria, D., Valley, C.T., Rodriguez-Martinez, J.A., Menichelli, E., Williamson, J.R., Ansari, A.Z., and Wickens, M. (2012). Cooperativity in RNA-Protein Interactions: Global Analysis of RNA Binding Specificity. Cell reports *1*, 570–581.

Caputi, M., and Zahler, A.M. (2001). Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family. The Journal of biological chemistry *276*, 43850–43859.

Chatrikhi, R., Wang, W., Gupta, A., Loerch, S., Maucuer, A., and Kielkopf, C.L. (2016). SF1 Phosphorylation Enhances Specific Binding to U2AF65 and Reduces Binding to 3′-Splice-Site RNA. Biophysical Journal *111*, 2570–2586.

Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nature reviews. Molecular cell biology *10*, 741–754.

Chhangawala, S., Rudy, G., Mason, C.E., and Rosenfeld, J.A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome biology *16*, 131.

Chung, H.-J., and Levens, D. (2005). c-myc expression: keep the noise down! Molecules and cells *20*, 157–166.

Cook, K.B., Hughes, T.R., and Morris, Q.D. (2014). High-throughput characterization of protein–RNA interactions. Briefings in Functional Genomics *14*, 74–89.

Coolidge, C.J., Seely, R.J., and Patton, J.G. (1997). Functional analysis of the polypyrimidine tract in pre-mRNA splicing. Nucleic acids research *25*, 888–896.

Corsini, L., Hothorn, M., Stier, G., Rybin, V., Scheffzek, K., Gibson, T.J., and Sattler, M. (2009). Dimerization and protein binding specificity of the U2AF homology motif of the splicing factor Puf60. The Journal of biological chemistry *284*, 630–639.

Darnell, R.B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley interdisciplinary reviews. RNA *1*, 266–286.

Dominguez, D., Tsai, Y.-H., Weatheritt, R., Wang, Y., Blencowe, B.J., and Wang, Z. (2016). An extensive program of periodic alternative splicing linked to cell cycle progression. eLife *5*.

Dredge, B.K., and Darnell, R.B. (2003). Nova regulates GABA(A) receptor gamma2 alternative splicing via a distal downstream UCAU-rich intronic splicing enhancer. Molecular and cellular biology *23*, 4687–4700.

Dredge, B.K., Stefani, G., Engelhard, C.C., and Darnell, R.B. (2005). Nova autoregulation reveals dual functions in neuronal splicing. The EMBO journal *24*, 1608–1620.

Dujardin, G., Lafaille, C., Petrillo, E., Buggiano, V., Gómez Acuña, L.I., Fiszbein, A., Godoy Herz, M.A., Nieto Moreno, N., Muñoz, M.J., Alló, M., Schor, I.E., and Kornblihtt, A.R. (2013). Transcriptional elongation and alternative splicing. Biochimica et biophysica acta *1829*, 134–140.

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 & development *8*, 465–480.

Ellington, A.D., and Szostak, J.W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature *346*, 818–822.

Fiszbein, A., and Kornblihtt, A.R. (2017). Alternative splicing switches: Important players in cell differentiation. BioEssays news and reviews in molecular, cellular and developmental biology *39*.

Fu, X.-D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nature reviews. Genetics *15*, 689–701.

Goren, A., Kim, E., Amit, M., Vaknin, K., Kfir, N., Ram, O., and Ast, G. (2010). Overlapping splicing regulatory motifs—combinatorial effects on splicing. Nucleic acids research *38*, 3318–3327.

Graveley, B.R., Hertel, K.J., and Maniatis, T. (2001). The role of U2AF35 and U2AF65 in enhancer-dependent splicing. RNA *7*, 806–818.

Han, H., Braunschweig, U., Gonatopoulos-Pournatzis, T., Weatheritt, R.J., Hirsch, C.L., Ha, K.C.H., Radovani, E., Nabeel-Shah, S., Sterne-Weiler, T., Wang, J., O'Hanlon, D., Pan, Q., Ray, D., Zheng, H., Vizeacoumar, F., Datti, A., Magomedova, L., Cummins, C.L., Hughes, T.R., Greenblatt, J.F., Wrana, J.L., Moffat, J., and Blencowe, B.J. (2017). Multilayered Control of Alternative Splicing Regulatory Networks by Transcription Factors. Molecular cell *65*, 539-553.e7.

Izquierdo, J.M. (2008). Hu antigen R (HuR) functions as an alternative pre-mRNA splicing regulator of Fas apoptosis-promoting receptor on exon definition. The Journal of biological chemistry *283*, 19077–19084.

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. The Journal of biological chemistry *289*, 17350–17364.

Jain, N., Lin, H.-C., Morgan, C.E., Harris, M.E., and Tolbert, B.S. (2017). Rules of RNA specificity of hnRNP A1 revealed by global and quantitative analysis of its affinity distribution. Proceedings of the National Academy of Sciences of the United States of America *114*, 2206–2211.

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. Nature structural & molecular biology *17*, 909–915.

Kralovicova, J., and Vorechovsky, I. (2017). Alternative splicing of U2AF1 reveals a shared repression mechanism for duplicated exons. Nucleic acids research *45*, 417–434.

Ladd, A.N., Nguyen, N.H., Malhotra, K., and Cooper, T.A. (2004). CELF6, a member of the CELF family of RNA-binding proteins, regulates muscle-specific splicing enhancer-dependent alternative splicing. The Journal of biological chemistry *279*, 17756–17764.

Lambert, D., and Draper, D.E. (2012). Denaturation of RNA secondary and tertiary structure by urea: simple unfolded state models and free energy parameters account for measured m-values. Biochemistry *51*, 9014–9026.

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P.A., and Burge, C.B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Molecular cell *54*, 887–900.

Lee, Y., and Rio, D.C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annual review of biochemistry *84*, 291–323.

Lehrach, H., Diamond, D., Wozney, J.M., and Boedtker, H. (1977). RNA molecular weight determinations by gel electrophoresis under denaturing conditions, a critical reexamination. Biochemistry *16*, 4743–4751.

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. Proceedings of the National Academy of Sciences of the United States of America *110*, E2687-95.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., Darnell, J.C., and Darnell, R.B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., Massirer, K.B., Pratt, G.A., Black, D.L., Gray, J.W., Conboy, J.G., and Yeo, G.W. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nature structural & molecular biology *20*, 1434–1442.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science (New York, N.Y.) *327*, 996–1000.

Luo, Z., Yang, Q., and Yang, L. (2016). RNA Structure Switches RBP Binding. Molecular cell *64*, 219–220.

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.

Marchese, D., Groot, N.S. de, Lorenzo Gotor, N., Livi, C.M., and Tartaglia, G.G. (2016). Advances in the characterization of RNA-binding proteins. Wiley interdisciplinary reviews. RNA *7*, 793–810.

Martin, L., Meier, M., Lyons, S.M., Sit, R.V., Marzluff, W.F., Quake, S.R., and Chang, H.Y. (2012). Systematic reconstruction of RNA functional motifs with high-throughput microfluidics. Nature methods *9*, 1192–1194.

Matera, A.G., and Wang, Z. (2014). A day in the life of the spliceosome. Nature reviews. Molecular cell biology *15*, 108–121.

Meng, Q., Rayala, S.K., Gururaj, A.E., Talukder, A.H., O'Malley, B.W., and Kumar, R. (2007). Signaling-dependent and coordinated regulation of transcription, splicing, and translation resides in a single coregulator, PCBP1. Proceedings of the National Academy of Sciences of the United States of America *104*, 5866–5871.

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 research *43*, 2378–2389.

Nagel, R.J., Lancaster, A.M., and Zahler, A.M. (1998). Specific binding of an exonic splicing enhancer by the pre-mRNA splicing factor SRp55. RNA *4*, 11–23.

Nettling, M., Treutler, H., Cerquides, J., and Grosse, I. (2016). Detecting and correcting the binding-affinity bias in ChIP-seq data using inter-species information. BMC genomics *17*, 347.

Oberstrass, F.C., Auweter, S.D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D.L., and Allain, F.H.-T. (2005). Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science (New York, N.Y.) *309*, 2054–2057.

Oltean, S., and Bates, D.O. (2014). Hallmarks of alternative splicing in cancer. Oncogene *33*, 5311–5318.

Papasaikas, P., and Valcárcel, J. (2016). The Spliceosome: The Ultimate RNA Chaperone and Sculptor. Trends in biochemical sciences *41*, 33–45.

Pascual, M., Vicente, M., Monferrer, L., and Artero, R. (2006). The Muscleblind family of proteins: an emerging class of regulators of developmentally programmed alternative splicing. Differentiation; research in biological diversity *74*, 65–80.

Rabenhorst, U., Beinoraviciute-Kellner, R., Brezniceanu, M.-L., Joos, S., Devens, F., Lichter, P., Rieker, R.J., Trojan, J., Chung, H.-J., Levens, D.L., and Zörnig, M. (2009). Overexpression of the far upstream element binding protein 1 in hepatocellular carcinoma is required for tumor growth. Hepatology (Baltimore, Md.) *50*, 1121–1129.

Ray, D., Kazan, H., Chan, E.T., Peña Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nature biotechnology *27*, 667–670.

Saulière, J., Sureau, A., Expert-Bezançon, A., and Marie, J. (2006). The Polypyrimidine Tract Binding Protein (PTB) Represses Splicing of Exon 6B from the β-Tropomyosin Pre-mRNA by Directly Interfering with the Binding of the U2AF65 Subunit ▿ . Molecular and cellular biology *26*, 8755–8769.

Schaub, M.C., Lopez, S.R., and Caputi, M. (2007). Members of the heterogeneous nuclear ribonucleoprotein H family activate splicing of an HIV-1 splicing substrate by promoting formation of ATP-dependent spliceosomal complexes. The Journal of biological chemistry *282*, 13617–13626.

Scotti, M.M., and Swanson, M.S. (2016). RNA mis-splicing in disease. Nature reviews. Genetics *17*, 19–32.

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. Molecular cell *11*, 965–976.

Shao, C., Yang, B., Wu, T., Huang, J., Tang, P., Zhou, Y., Zhou, J., Qiu, J., Jiang, L., Li, H., Chen, G., Sun, H., Zhang, Y., Denise, A., Zhang, D.-E., and Fu, X.-D. (2014). Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. Nature structural & molecular biology *21*, 997–1005.

Shi, H., Hoffman, B.E., and Lis, J.T. (1997). A specific RNA hairpin loop structure binds the RNA recognition motifs of the Drosophila SR protein B52. Molecular and cellular biology *17*, 2649–2657.

Shi, Y. (2017). The Spliceosome: A Protein-Directed Metalloribozyme. Journal of molecular biology *429*, 2640–2653.

Sickmier, E.A., Frato, K.E., Shen, H., Paranawithana, S.R., Green, M.R., and Kielkopf, C.L. (2006). Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. Molecular cell *23*, 49–59.

Singh, N.N., Singh, R.N., and Androphy, E.J. (2007). Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. Nucleic acids research *35*, 371–389.

Singh, R., Valcárcel, J., and Green, M.R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science (New York, N.Y.) *268*, 1173–1176.

Sirand-Pugnet, P., Durosay, P., Clouet d'Orval, B.C., Brody, E., and Marie, J. (1995). beta-Tropomyosin pre-mRNA folding around a muscle-specific exon interferes with several steps of spliceosome assembly. Journal of molecular biology *251*, 591–602.

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 (New York, N.Y.) *312*, 1961–1965.

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 biology *13*, R67.

Sutandy, F.X.R., Ebersberger, S., Huang, L., Busch, A., Bach, M., Kang, H.S., Fallmann, J., Maticzka, D., Backofen, R., Stadler, P.F., Zarnack, K., Sattler, M., Legewie, S., König, J. (2018). In vitro iCLIP-based modeling uncovers how the splicing factor U2AF2 relies on regulation by cofactors. Genome research *28*, 699-713

Sutandy, F.X.R., Hildebrandt, A., and König, J. (2016). Profiling the Binding Sites of RNA-Binding Proteins with Nucleotide Resolution Using iCLIP. Methods in molecular biology (Clifton, N.J.) *1358*, 175–195.

Tavanez, J.P., Madl, T., Kooshapur, H., Sattler, M., and Valcárcel, J. (2012). hnRNP A1 proofreads 3' splice site recognition by U2AF. Molecular cell *45*, 314–329.

Tripathi, V., Sixt, K.M., Gao, S., Xu, X., Huang, J., Weigert, R., Zhou, M., and Zhang, Y.E. (2016). Direct Regulation of Alternative Splicing by SMAD3 through PCBP1 Is Essential to the Tumor-Promoting Role of TGF-β. Molecular cell *64*, 549–564.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science (New York, N.Y.) *302*, 1212–1215.

Urlaub, H., Hartmuth, K., and Lührmann, R. (2002). A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. Methods (San Diego, Calif.) *26*, 170–181.

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 (New York, N.Y.) *273*, 1706–1709.

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. Proceedings of the National Academy of Sciences of the United States of America *113*, E7169-E7175.

Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D.L., Nutter, R.C., Segal, E., and Chang, H.Y. (2012). Genome-wide measurement of RNA folding energies. Molecular cell *48*, 169–181.

Warf, M.B., and Berglund, J.A. (2010). Role of RNA structure in regulating pre-mRNA splicing. Trends in biochemical sciences *35*, 169–178.

Warf, M.B., Diegel, J.V., Hippel, P.H. von, and Berglund, J.A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proceedings of the National Academy of Sciences of the United States of America *106*, 9203–9208.

Will, C.L., and Lührmann, R. (2011). Spliceosome structure and function. Cold Spring Harbor perspectives in biology *3*.

Wu, J.Y., and Maniatis, T. (1993). Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. Cell *75*, 1061–1070.

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.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D.L., Sun, H., Fu, X.-D., and Zhang, Y. (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Molecular cell *36*, 996–1006.

Yardımcı, G.G., Frank, C.L., Crawford, G.E., and Ohler, U. (2014). Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic acids research *42*, 11865–11878.

Yi, J., Shen, H.-F., Qiu, J.-S., Huang, M.-F., Zhang, W.-J., Ding, J.-C., Zhu, X.-Y., Zhou, Y., Fu, X.-D., and Liu, W. (2016). JMJD6 and U2AF65 co-regulate alternative splicing in both JMJD6 enzymatic activity dependent and independent manner. Nucleic acids research *45*, 3503–3518.

Zamore, P.D., Patton, J.G., and Green, M.R. (1992). Cloning and domain structure of the mammalian splicing factor U2AF. Nature *355*, 609–614.

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.

Zuo, P., and Maniatis, T. (1996). The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. Genes & development *10*, 1356–1368.

7. APPENDIX

7.1 List of Figures

Figure 1. Stages of splicing in eukaryotes 5
Figure 2. Pictogram of the U2-dependent splice site recognition elements 7
Figure 3. Schematic of the splicing regulatory events 8
Figure 4. Domain architecture and conformation of U2AF65 12
Figure 5. Overview of iCLIP protocol to study RNA-protein interactions 15
Figure 6. *In vitro* iCLIP protocol 44
Figure 7. Optimization of *in vitro* iCLIP library preparation 47
Figure 8. Initial inspection of U2AF65^{RRM12} *in vitro* iCLIP binding landscape 49
Figure 9. Exploration of U2AF65 *in vivo-in vitro* iCLIP landscapes 50
Figure 10. Mathematical modeling of U2AF65 binding site affinities (K_d values) from *in vitro* iCLIP titration assays 52
Figure 11. Validation of K_d values measured by *in vitro* iCLIP 53
Figure 12. Motif and distribution of binding site affinities measured by *in vitro* iCLIP 55
Figure 13. *In vivo – in vitro* iCLIP comparative modeling 57
Figure 14. U2AF65 interactions with U2AF35 and SF1 only partially explain 3′ splice site stabilization *in vivo* 58
Figure 15. *In vivo* modulation of U2AF65 binding upon partial U2AF65 knockdown 59
Figure 16. Identification of U2AF65 regulators with Random Forests 60
Figure 17. Co-factors change U2AF65RRM12 binding *in vitro* 63
Figure 18. Recapitulation of regulation by hnRNPC *in vitro* 65
Figure 19. *In vitro* experiment captures in vivo regulation at PTBP2 exon 10 67
Figure 20. Minigene reporter assays confirm PTBP2 exon 10 regulation by PTBP1 and FUBP1 *in vivo* 69
Figure 21. *In vitro* changes predict *in vivo* regulation of alternative splicing 71
Figure 22. Predicted regulator RBPs affects *in vivo* alternative splicing decision 72
Figure 23. Working model of U2AF65 recruitment to 3′ splice site 78

7.2 List of Tables

Table 1. List of *in vitro* transcripts used in *in vitro* iCLIP experiments. 45
Table 2. RNA oligos that were used in K_d measurement with MST and ITC. 54
Table 3. Mutations list for minigenes construct of *PTBP2* exon 10 splicing assay. 66

7.3 Abbreviations

BCA	bicinchoninic acid
BSA	bovine serum albumin
cDNA	complementary deoxyribonucleic acid
Chr	chromosome
CTD	carboxy terminal domain
Cy5	cyanine5
C4BPB	complement component 4 binding protein beta
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
ELAVL	ELAV-like protein
FUBP1	far upstream element binding protein 1
HRP	horseradish peroxidase
IGV	Integrative Genomics Viewer
IMB	Institute of Molecular Biology
IPTG	isopropyl β-D-1-thiogalactopyranoside
K	kilo
KHDRBS1	KH RNA binding domain containing, signal transduction associated 1
M	molar
MALAT1	metastasis associated lung adenocarcinoma transcript 1
MBNL1	muscleblind-like 1
MOPS	3-(N-morpholino)propanesulfonic acid
MAT2A	methionine adenosyltransferase 2A
mRNA	messenger ribonucleic acid
mRNP	messenger ribonucleic acid-protein complex
MYL6	myosin light chain 6
NF1	neurofibromin 1
NOVA1	neuro-oncological ventral antigen 1

nt	nucleotide
NUP133	nucleoporin 133kDa
PAPD4	poly(A) RNA polymerase D4
PCBP	poly r(C) binding protein
PCR	Polymerase Chain Reaction
PEG	polyethylene glycol
PK	proteinase K
PNK	polynucleotide kinase
Pol II	polymerase II
PTBP	polypyrimidine tract-binding protein
qPCR	quantitative polymerase chain reaction
RBM	RNA binding motif protein
RNA	ribonucleic acid
RS	arginine and serine residues
RT	reverse transcritption
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SF1	splicing factor 1
SF3B155	splicing factor 3b, subunit 1, 155kDa
siRNA	small interfering RNA
snRNA	small nuclear RNA
SNRPA	small nuclear ribonucleoprotein polypeptide A
SRp55	splicing factor, arginine/serine-rich, 55 KDa
TBE	Tris/Borate/EDTA
UV	ultraviolet
^{o}C	degree Celcius

8. CURRICULUM VITAE

Personal Information
Name: Reymond Sutandy
Date of birth: 13 December 1988
Place of birth: Klaten
Nationality: Indonesia
Email: r.sutandy@imb-mainz.de

Education

2013-2018
PhD., Institute of Molecular Biology (IMB), JGU Mainz, Germany,
Supervised by Dr. Julian König
PhD Thesis:
Deciphering the binding regulation of the core splicing factor U2AF65 using *in vitro* iCLIP

2011-2013
MSc., Graduate School of Systems Biology and Bioinformatics, National Central University, Taiwan
Supervised by Dr. Chien-Sheng Chen
Master Thesis:
Heterogeneous ribonucleoprotein K (hnRNP K) inhibits the post-transcriptional regulation of mature miRNA 122

2007-2011
BSc., School of Life Science and Technology, Bandung Institute of Tecnology, Indonesia
Supervised by Dr. Ernawati Arifin Giri Rachman
Bachelor Thesis:
Transformation and analysis of L-HbsAg DNA transient expression in tobacco (*Nicotiana tabacum* Linn.) seedlings